Insights into hominid evolution from the gorilla genome sequence

Supplementary Information

Aylwyn Scally *et al.*

External supplementary tables

Some supplementary tables referred to in this document are found in separate Supplementary files:

- ST3.1: in Supplementary file 2
- ST4.3: in Supplementary file 6
- ST4.5: in Supplementary file 7
- ST8.1: in Supplementary file 4
- ST8.2: in Supplementary file 2
- ST8.3: in Supplementary file 2
- ST8.4: in Supplementary file 2
- ST8.5: in Supplementary file 5
- ST8.7: in Supplementary file 2
- ST8.8: in Supplementary file 2
- ST8.9: in Supplementary file 2
- ST11.1: in Supplementary file 2
- ST12.2: in Supplementary file 3

1. Taxonomy of the extant great apes

Family	Hominidae (hominids; extant great apes)						
Subfamily	Homininae (hominines; extant African great apes)					Ponginae (pongines)	
Tribe	Hominini (hominins)	Panini (panins)		Gorillini (gorillins)		Pongini (pongins)	
Genus	*Homo*	*Pan*		*Gorilla*		*Pongo*	
Species	***H. sapiens*** (modern human)	***P. troglodytes*** (chimpanzee)	*P. paniscus* (bonobo)	***G. gorilla*** (Western gorilla)	*G. beringei* (Eastern gorilla)	*P. pygmaeus* (Bornean orangutan)	***P. abelii*** (Sumatran orangutan)

Table ST1.1. Taxonomy of the extant great ape family, following Wood and Harrison (Wood and Harrison 2011). Publicly available reference genomes exist for species in boldface. (*Pongo abelii* was formerly designated *P. pygmaeus abelii*, a subspecies of *P. pygmaeus*, but is now accepted as a separate species (Singleton et al. 2008).)

2. Genome Sequencing

	Kamilah	Kwanza	Mukisi	EB(JC)
Species	*G. gorilla gorilla* (Western Lowland)	*G. gorilla gorilla* (Western Lowland)	*G. beringei graueri* (Eastern Lowland)	*G. gorilla gorilla* (Western Lowland)
Sex	Female	Male	Male	Female
Date of birth	5 Dec 1977	2 Mar 1989	1 Jan 1957	Unavailable
Location	San Diego Wild Animal Park, San Diego, California USA	Lincoln Park Zoo, Chicago, Illinois USA	(Died 17 Dec 2000)	Unavailable
Stud number	0661	9912	1107	Unavailable

Table ST2.1: Gorillas sampled in this study

Assembly data

The primary sequence data for assembly comprised 5.4 Gbp of ABI capillary-sequenced whole-genome shotgun read pairs, and 166.1 Gbp of Illumina short read pairs, mostly 35-37 bp in length with library insert sizes of 130-450 bp (Table ST2.2, Table ST2.3). All sequence was derived from DNA sampled from a single female Western Lowland gorilla (*Gorilla gorilla gorilla*), Kamilah (Table ST2.1). All data were collected using standard sequencing protocols.

Library SRA accession number(s)	Mean insert size (bp)	Read length (bp)	Read pairs sequenced	Total Mbp sequenced
ERX000198	130	37	193,010,179	14,282
ERX000194-ERX000196	150	34-37	610,104,905	43,604
ERX000200	160	35	397,082,523	27,795
ERX000199	170	35	441,573,095	30,910
ERX007927	420	36	3,019,266	217
ERX000190-ERX000191	450	36-37	44,473,981	3209
ERX000192, ERX000723-ERX000726, ERX000728, ERX000729, ERX000731	450	76	300,843,751	45,728
ERX000193	3000	37	6,381,041	472

Table ST2.2: Whole genome Illumina sequence data for Kamilah (SRA sample accession number ERS000016).

NCBI Trace Archive CENTER_PROJECT	Insert range (kbp)	Reads	Read length (bp)
GORILLA-WGS-92433	2-3	2743	713
GORILLA-WGS-92434	2.5-3	355,483	735
GORILLA-WGS-92435	3-3.5	4218	745
GORILLA-WGS-92436	3-3.5	4408	782
GORILLA-WGS-92437	3.5-4	1605	625
GORILLA-WGS-92438	3.5-4	248,658	646
GORILLA-WGS-92439	4-5	303,819	746
GORILLA-WGS-92440	4-5	5,040,566	745
GORILLA-WGS-92441	5-6	1,518,041	635
GORILLA-WGS-92442	6-7	9541	624
GORILLA-WGS-92443	7-8	3979	610
GORILLA-WGS-92444	8-10	2421	606
GORILLA-WGS-92445	7-10	1960	579
GORILLA-WGS-92446	8-12	2552	615
		Total reads	7,499,994
		Total bp	5,389,840,322

Table ST2.3: Whole genome ABI capillary sequence data for Kamilah

Other gorilla data

To explore genetic variation within the *Gorilla* genus, we also collected sequence data from two further Western Lowland gorillas, Kwan and EB(JC), and an Eastern Lowland gorilla, Mukisi (Table ST2.1). Low coverage whole genome sequence data was obtained for Mukisi and Kwan (Table ST2.4). In addition, to obtain high depth for variant calling at selected sites genome wide, we created reduced representation libraries for EB(JC) and Mukisi. The same representation (restriction enzyme and gel slice selection) was used for both libraries, and both samples were multiplexed on a single lane of the sequencing machine (Table ST2.4).

Individual	Strategy	SRA accession number	Mean insert size (bp)	Read length (bp)	Read pairs sequenced	Total Mbp sequenced
Mukisi	WG	ERS004138	190	37-54	387,932,732	38,017
Mukisi	RR	ERS008713	220	95	61,732,672	11,729
EB(JC)	RR	ERS008712	190	95	51,236,872	9,735
Kwanza	WG	SRX023771	320	37	68,062,949	5,036
Kwanza	WG	SRX023772	320	35	1,186,048,916	83,023
Kwanza	WG	SRX023773	310	35	318,062,772	22,264

Table ST2.4: Whole genome (WG) and reduced representation (RR) Illumina libraries for Mukisi, EB(JC) and Kwanza

Chimpanzee and bonobo data

We also collected low coverage whole genome sequence data from a female chimpanzee (*Pan troglodytes*) 'chimp1' and a male bonobo (*Pan paniscus*) 'bonobo1' (Table ST2.5).

Individual	Strategy	SRA accession number	Mean insert size (bp)	Read length (bp)	Read pairs sequenced	Total Mbp sequenced
chimp1	WG	ERS002007	400	76	16,496,692	2,507
bonobo1	WG	ERS002008	300	76	59,310,478	9,015

Table ST2.5: Whole genome Illumina libraries for chimp1 and bonobo1.

3. Assembly

Introduction

Recent developments in sequencing technology have transformed the field of genome assembly and analysis. In comparison with the preceding technology of capillary sequencing, the current generation of high-throughput machines has reduced the costs and timescales of raw sequence production by several orders of magnitude. The much higher attainable yields mean that in human and other organisms with moderate genomic GC content, over 99% of the sequenceable genome is sampled (Bentley et al. 2008), and the vast majority of nucleotides are read many times, mitigating the effects of sequencing errors.

However, the new technologies still suffer from an important disadvantage with respect to capillary sequencing, which is that the read lengths they produce are much shorter (35-100 bp compared to 600-700 bp). This has a significant effect on the difficulty of assembly, and the initial assemblies produced from short reads tend to have lower contiguity (assemblies are very fragmented) and less accurate representation of repetitive sequence structure than those derived from capillary reads. These problems are particularly relevant for vertebrate genomes, which have comparatively high repeat content, and the assembly of such genomes therefore remains a difficult computational problem. Furthermore the quantity of data obtainable with high-throughput sequencing also brings with it much greater informatics challenges of processing and storage.

For the gorilla assembly we were able to combine both capillary and next generation sequence data, thereby benefitting from the contiguity of the former and the high coverage of the latter. We also took advantage of ordering information (though not sequence) from the human assembly to guide and improve the long-range and chromosomal structure of the assembly. Below we describe in detail the assembly pipeline and methods used, followed by an assessment of assembly quality and accuracy.

Assembly method

The assembly pipeline involved several phases, and incorporated a variety of tools for low-level sequence assembly and alignment.

1. To begin with, a *de novo* assembly of the Illumina data was produced using the ABySS assembler (Simpson, Wong et al. 2009). Only the first stage of ABySS was used, assembling the reads themselves but not making use of pairing information. This produced 5,624,115 contigs with an N50 of 546 bp. Contigs longer than 50 bp were then added to the collection of WGS capillary read pairs and a further assembly was produced using the Phusion assembler (Mullikin and Ning 2003). This 'seed' assembly comprised 1,179,807 contigs with an N50 of 2867 bp. Using the pairing information in the WGS reads, Phusion was able to place two thirds of these contigs into supercontigs having an N50 of 15,739 bp.

In the subsequent phases, the aim was to grow these seed contigs and where possible merge them using Illumina read pairs from the initial data. To prepare for this, an alignment of all the Illumina data to the seed assembly was produced using Maq (Li, Ruan et al. 2008).

2. To improve the long-range structure of the assembly, supercontig construction was guided by human homology. SSAHA (Ning, Cox et al. 2001) was used to map the seed contigs to the human genome, and was able to place 96% of them. Some contigs (representing repetitive sequence) placed in multiple locations, and some placed in locations that were fully spanned by other contigs. The latter placings were removed, but not the former, on the grounds that Phusion may have over-collapsed repetitive sequence.

Supercontigs were then constructed by taking seed contigs in order of their placing on human (including duplicate placings), stopping wherever a potential break between human and gorilla was inferred. In principle, a break was inferred at any position at which there were no spanning read pairs in an alignment of all the Illumina data to human. (Maq was used to produce such an alignment.) These occurred approximately every 100 kbp. However, if the placing of seed contigs around that position was consistent with their ordering and spacing in the original de novo supercontigs made by Phusion, so that the break was effectively spanned by part of a de novo supercontig, then that break was ignored. A break was also ignored if it was directly spanned by a seed contig. Since over 90% of the human genome was spanned in this way, most breaks in the Illumina alignment did not break supercontig construction, and the resulting supercontig N50 was over 900 kbp.

Supercontigs constructed from placed seed contigs are referred to as 'placed' supercontigs. The 4% of seed contigs which did not place on human were passed to the next phase as 'unplaced' supercontigs.

3. A set of local assemblies was constructed, one for each supercontig, comprising the seed contigs in that supercontig and all the Illumina reads mapping to them, plus unmapped mates. (Reads were allowed to be included in more than one assembly.) Within the placed supercontigs the homology with human helped, since it was possible to also include Illumina reads mapping (with Maq) to the homologous region of human (and their unmapped mates). For computational efficiency, local assemblies were divided into sub-assemblies of length 50 kbp, and subsequently spliced back together after completion. Sub-assemblies were carried out in parallel on a multi-core compute farm.

4. Each sub-assembly proceeded as follows. Where Illumina reads had been drawn both from alignments to the seed assembly and human (i.e. in placed supercontigs), a unique set was constructed by removing duplicates. Then, assembly of this set plus the seed contigs was carried out using Velvet (Zerbino and Birney 2008). In each case the Illumina reads were grouped by library, and the insert size distribution parameters of each library were provided along with kmer size and coverage parameters required by Velvet. The seed contigs were supplied as ‘long' sequences, to be used by Velvet in resolving ambiguities in its assembly graph.

The final graph produced by Velvet, with nodes representing stretches of assembled sequence and edges specifying alternative connections between them, was then used to construct the new supercontig. First the seed contigs were mapped onto this graph. Then, taking their order within the original supercontig, a path was sought within the graph from each contig to the next, the search extending out to twenty nodes. Where such a path existed, a contig and its successor were replaced by the sequence corresponding to the nodes making up the path. If the path was of zero length (i.e. the contigs mapped to the same node), this was called a merge, and often multiple contigs were merged in this way. Longer paths were called path joins. Where multiple paths existed, the shortest was taken. Where no path was found, the contig was merely extended (if possible) by sequence from the nodes onto which its ends had been mapped. The outcome of this process was a new scaffold in which, in general, most of the original seed contigs had been extended and many had been merged with their neighbours. Gaps between the new contigs were estimated from the original separation of seed contigs on human, taking into account the amount of sequence extension during assembly, and with negative estimates replaced by a gap of 10 bp.

In some cases it was found that repetitive sequence within the seed contigs gave a greater risk of misassembly. A parameter counting the number of duplicate node occurrences was used to characterise this.

This whole procedure (Velvet assembly followed by meta-assembly) was repeated three times, once for each of the kmer sizes 21, 23 and 25 - different kmer sizes were found to be effective in different regions. The resulting assemblies were ranked by number of merges, number of path joins, the misassembly parameter, and total amount of sequence added, and only the highest ranking one kept.

Carrying this out across the genome increased the contig N50 of the whole assembly from 2.9 to 11.9 kbp, and increased genomic representation to 95%.

5. The first postprocessing phase involved concatenating the sub-assemblies within supercontigs to make new whole supercontigs. Consecutive sub-assemblies were checked for overlapping sequence at adjacent ends, and spliced together where possible. In some cases this splicing was possible only if some sequence was cut from the ends of one or both of the sub-assemblies, and this is what was done if the amount of sequence to be cut was less than half the length of overlap. The resulting cut fragments were kept as separate contigs in the assembly if they were 100 bp or longer. (Evidence from remapped read depth suggests that a substantial fraction of these cut fragments come from heterozygous structures, representing sequences present in only one of the two diploid chromosomes) Where splicing was not possible, a gap was inserted, estimated as above from the initial separation of the sub-assemblies on human and the amount of sequence extension at each end, with negative estimates replaced by 10 bp.

6. The second postprocessing phase involved error correction using the consensus of Illumina reads aligned to the assembly, intended primarily to remove base substitutions and short indels derived from read sequencing errors (which were expected to be prevalent in stretches of sequence derived from the low coverage capillary reads). As before, this alignment was done using Maq, and then a consensus including both homozygous and heterozygous variants (SNPs and indels) was called. These variants were filtered using the SAMtools program (Li et al. 2009) with a minimum depth of ten reads. This left 397,237 homozygous SNPs and 79,731 homozygous indels, all of which were assumed to be errors in the assembly, were used to correct it.

Chromosomes comprising placed supercontigs were constructed, incorporating the human 5/17 reciprocal translocation and the separation of regions corresponding to human chromosome 2 into gorilla chromosomes 2a and 2b, using the McConkey (McConkey 2004) nomenclature. The completed assembly comprises 2.9 Gbp on 24 chromosomes with 117 Mbp of unplaced material

The version of the assembly at this stage was used for the majority of analysis, and was released in Ensembl as gorGor3.

Fig. SF3.1 illustrates the flow of sequence and information through the assembly pipeline up to this point.

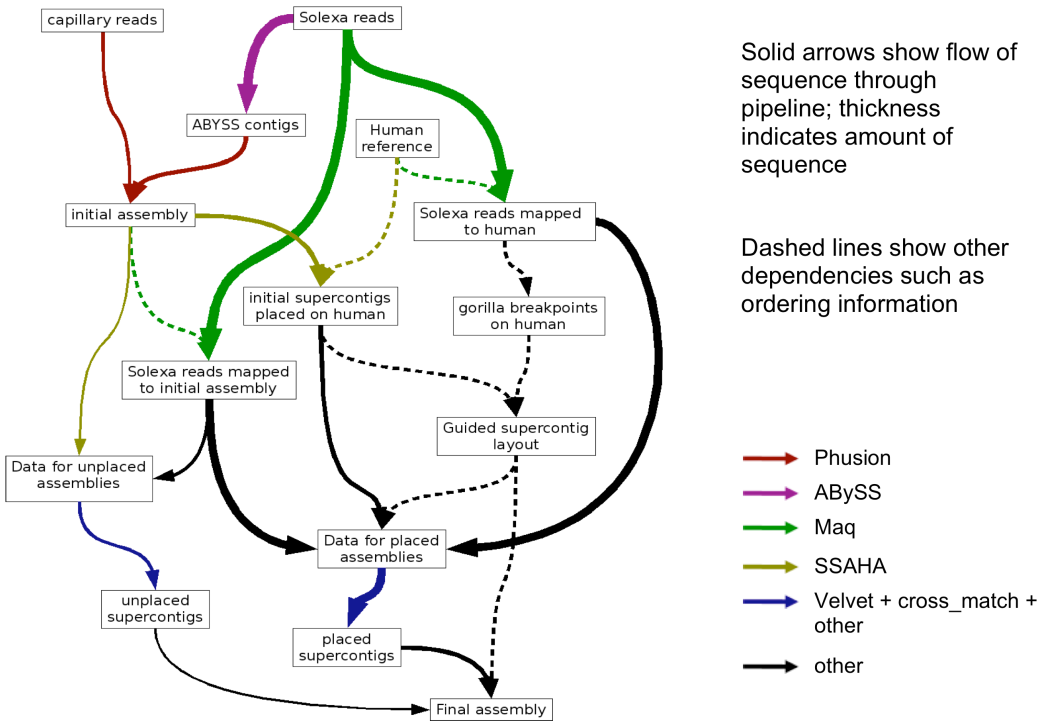

capillary reads
Solexa reads
ABYSS contigs
Human reference
initial assembly
Solexa reads mapped to human
initial supercontigs placed on human
gorilla breakpoints on human
Solexa reads mapped to initial assembly
Guided supercontig layout
Data for unplaced assemblies
Data for placed assemblies
unplaced supercontigs
placed supercontigs
Final assembly
Solid arrows show flow of sequence through pipeline; thickness indicates amount of sequence
Dashed lines show other dependencies such as ordering information
Phusion
ABySS
Maq
SSAHA
Velvet + cross_match + other
other

Figure SF3.1 Assembly pipeline showing the tools used and the flow of sequence (capillary reads and Solexa (Illumina) reads) and ordering information.

7. The final assembly phase made further improvements to long range structure, up to and including the chromosomal scale. Additional Kamilah paired-end sequence data obtained from BAC and fosmid clones at 11.3x and 1.4x spanning coverage respectively was used to help order and orient scaffolds into chromosomes, taking the order implied by human homology wherever there was no contrary evidence from clone ends.

To create the chromosomal AGP files, the assembly data were aligned against the human genome with LASTZ to align and score non-repetitive gorilla regions against repeat-masked human sequence. Alignment chains were differentiated between orthologous and paralogous alignments and only "reciprocal best" alignments were retained in the alignment set. The gorilla AGP files were generated from these alignments in a manner similar to that already described (CGSAC 2005).

A total of 21 breaks (list of breakpoints provided below) based on documented human/gorilla differences (Stanyon, Rocchi et al. 2008) were introduced.

1. human HSA2 fusion, with HSA2:0Mb-111.9Mb on GGO2A and HSA2:111.9-242Mb on GGO2B. There is an inversion with GGO2A HSA2:94-111.9Mb followed by HSA2:90Mb-0Mb.
2. GGO4, an inversion of the region corresponding to HSA4:49.3-70.0Mb
3. the GGO-specific HSA5/17 reciprocal translocation:
   GGO5 is HSA17:78.5-15.4Mb followed by HSA5:79.9-180.7Mb
   GGO17 is HSA17:0-15.4Mb followed by HSA5:80-0Mb
4. GGO7, an inversion of the region corresponding to HSA7:76.5-102.0Mb
5. GGO8, an inversion of the region corresponding to HSA8:30.0-86.9Mb
6. GGO9, an inversion of the region corresponding to HSA9:0Mb-70.0Mb
7. GGO10, an inversion of the region corresponding to HSA10:27.6-80.9Mb
8. GGO12, an inversion of the region corresponding to HSA12:21.2-63.6Mb
9. GGO14, an inversion of the region corresponding to HSA14:0-44.8Mb
10. GGO18, an inversion of the region corresponding to HSA18:0-14.9Mb

Gap sizes were estimated based on the human genome where possible. Centromeres were introduced into the gorilla sequence at the positions of the centromeres in the human chromosomes. On chromosome GGO2B (HSA2q), the centromere was inserted at the location of the ancestral centromere localized specifically by identifying its likely location by the presence of alpha-satellite repetitive elements.

Alignments to human mRNAs revealed 438 "joins" between neighbouring contigs. In all but one case, the chromosomal AGP files were consistent with the order/orientation suggested by the human mRNAs. In one case, the human mRNA was interrupted by a documented human/gorilla inversion supported by several BAC clones.

Finally, unplaced contigs shorter than 2kbp were removed. The completed assembly comprises 2.9 Gbp ordered and oriented along 24 chromosomes with 170 Mbp of unplaced material This version will be released as gorGor4.

Assembly quality

Assembly quality was assessed in two principal ways: alignment of Illumina data to the assembly and direct comparison with a set of finished whole fosmid sequences from the same individual.

1. The Illumina data was mapped to the assembly using Maq; in total 94% of reads mapped. As in the error correction phase of assembly, homozygous and heterozygous variants were called from this alignment, this time finding 106,407 homozygous SNPs and 10,412 homozygous indels in total (all of which are likely to be assembly errors). The counts for heterozygous SNPs and indels were 4,943,562 and 597,683 respectively.

2. Twelve clones were selected at random from a fosmid library for Kamilah, and finished sequencing of them attempted, which was successful on all but two. The resulting ten sequences, of length 30-40 kbp, were mapped to the assembly using bwa (Li and Durbin 2010). The alignments were postprocessed to identify differences (at single bases and indels as well as larger rearrangements) between the finished sequences and the assembly.

Seven of the ten fosmids mapped with more than 92% of sequence on one supercontig. Of the remaining three, one mapped with 88% on one supercontig, one with 84%, and one with 98% on two supercontigs.

Within the aligned blocks, single-base differences and indels were taken to be assembly errors, except where they coincided with heterozygous positions identified in the alignment of Illumina data or uncertainties in the finished fosmid sequences. Such errors were overwhelmingly found within a small number of short regions, each less than a few hundred bp in length. Errors were sufficiently dense within these regions that the median error separation was just 4 bp, but across the whole alignment stretches of error-free sequence were typically 7.2 kbp as measured by the N50 length. Error clusters tended to be found in repetitive regions of the genome or close to contig ends. In the 40% of the assembly which was outside RepeatMasked regions and no closer than 50 bp to a contig end, the N50 error-free length was greater than 10 kbp and even the total rate of single-base and indel errors was just 0.13 per kbp.

At larger scales, the alignment included 23 gaps in the assembly larger than 200 bp, of which 21 had been estimated to within plus or minus 30% of their correct length. One large misassembly was found: a 5 kbp segment inverted and transposed by 11.5 kbp. (A check revealed no such corresponding rearrangement in the human genome, indicating that it was not an error caused by reliance on human homology.)

An alternative assembly-wide analysis, based on an alignment of the human, chimpanzee and gorilla assemblies, derived an indel-based metric by comparing the observed number of gap-delimited segments in orthologous aligned sequence with the expectation under a model of indel evolution (Meader, Hillier et al. 2010). Assigning indels to each of the three lineages by parsimony, we obtained a value on human of 0.103 per kbp (95% c.i. 0.097 - 0.109), for chimpanzee 0.231 per kbp (95% c.i. 0.225-0.237), and for gorilla 0.236 per kbp (95% c.i. 0.229-0.242), suggesting that the gorilla assembly is comparable to the chimpanzee assembly. according to this metric, about 2-2.5 times as bad as the human reference assembly.

Annotation

Annotation of genes and transcripts across the whole assembly was carried out using the Ensembl pipeline, drawing on evidence projected from human genome annotation and from alignment of human, gorilla and other vertebrate proteins from the Ensembl and Uniprot databases. Predictions of gene orthologues and paralogues between and within species were carried out using the Ensembl-Compara pipeline (Vilella et al. 2009).

lincRNA

Long intergenic RNA (lincRNA) transcript models were obtained from the ENSEMBL pipeline through filtering. Excluded were transcript models: (1) overlapping the exons or introns of protein coding genes (ENSEMBL) and the RefSeq (Pruitt, Tatusova et al. 2005) track from UCSC genome browser (UCSC (Kent 2002) mapped from human to gorilla), (2) transcript models of less than 200 bp length and (3) those with predicted coding potential (Kong, Zhang et al. 2007).

Constraint of a lincRNA transcript was defined as the quotient of the nucleotide substitution rate of the transcript and the average nucleotide substitution rate of ancestral repeats within 20kb of the lincRNA transcripts. Ancestral repeats and rates were derived using the rhesus macaque (rheMac2) genome assembly. Nucleotide substitution rates were estimated using the general reversible model implemented in the PAML4 package (Yang 2007). We employed the Wilcoxon rank sum test to assess the constraint of lincRNAs compared to ancestral repeats (median 12% suppression; $p<10^{-5}$). (Table ST3.1)

Multiple primate whole genome alignment

We included the gorilla assembly with human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), orangutan (*Pongo abelii*) and macaque (*Macaca mulatta*) in a 5-way whole genome alignment using the Enredo-Pecan-Ortheus (EPO) pipeline (Paten et al. 2008). The alignment as a whole contains 80-90% of each reference genome (Table ST3.2), and raw nucleotide divergence rates calculated within aligned blocks (Table ST3.3) are consistent with those observed in previous studies.

reference genome	percentage included
gorilla	90.6
human	89.5
chimpanzee	83.1
orangutan	79.5
macaque	85.8

Table ST3.2: Percentage of each reference genome included in the EPO 5-way alignment.

	human	chimpanzee	orangutan	macaque
gorilla	1.75	1.81	3.50	6.29
human	-	1.37	3.40	6.23
chimpanzee	1.81	-	3.44	6.27
orangutan	3.40	3.44	-	6.35

Table ST3.3: Percentage mismatch rates between species in the EPO 5-way alignment.

4. CoalHMM and great ape speciation

Pre-processing of the data

We filtered the 5-species EPO alignments to remove low quality assembly and ambiguous alignment regions. First we removed all the synteny blocks containing zero or multiple sequences for any of human, chimpanzee, gorilla and orangutan. We also removed all positions having a gap in any of the three species human, chimpanzee and gorilla.

The model of sequence evolution underlying CoalHMM makes multiple assumptions. First it assumes a molecular clock. Violation of this assumption in ingroup species (human, chimpanzee and gorilla) can lead to biases in the estimates of ILS, notably by introducing asymmetry between frequency of the HG,C and CG,H topologies (see Hobolth et al. (2011), supplementary material 1, but also Slatkin (2008)). A simple phylogenetic analysis shows that in the raw data the chimpanzee branch length is longer than the human branch, probably due to a higher sequence error rate on this genome (Hobolth et al. 2011).

To identify regions with low quality score in the chimpanzee assembly, we used a 10-nucleotide sliding window, and computed the average sequence quality score in the window. All windows with an average score lower than a threshold parameter were discarded, and the synteny blocks split accordingly. (This step was skipped for chromosome 21, where a higher-quality chimpanzee sequence was available.) The speciation analysis was repeated with various threshold values, and a value of 7 was found to be the lowest for which the number of positions sorting as ((H,G),C) equalled the number sorting as ((C,G),H). This threshold also proved to be sufficient to remove the departure from a molecular clock, as assessed by comparing pairwise HG and CG divergences over windows of 1Mb. Nevertheless to further remove ambiguously aligned regions we used a 50-nucleotide sliding window and removed windows with at least 50 gaps summed across all species. The resulting data set, which we call 'FULL', covers 2,006,962,657 positions of the alignment.

To test the effect of repetitive material on our results, we applied a further filter using a 5-nucleotide sliding window and removed regions with at least 5 sites masked by RepeatMasker summed across all species. The resulting data set covers 654,272,926 positions of the full alignment. This second dataset, which we call 'MASKED', displays a higher proportion of conserved regions and contains 20% only of the human genome.

CoalHMM analysis

We gathered the resulting blocks into chunks of ~1Mb each, and analyzed them independently. We used the CoalHMM software with the model introduced in Hobolth et al 2007 (Hobolth, Christensen et al. 2007), and further developed in Dutheil et al (Dutheil, Ganapathy et al. 2009). This method uses a hidden Markov chain to model the transitions between genealogies along the genome alignment. We reset the Markov chain for each block (that is, when there is a synteny break or when there is a gap in the alignment, because some regions have been removed by the filtering), and estimated one set of parameters per 1Mb chunk. A posterior decoding approach was used to reconstruct the most likely genealogy for each position in the genome (Dutheil, Ganapathy et al. 2009).

We used the bias correction procedure described in Dutheil et al (Dutheil, Ganapathy et al. 2009), using data simulated with parameter values given in Table ST4.1.

$N_e[H] = N_e[C] = N_e[G]$	30,000
$N_e[HC]$	(30,000; 60,000; 120,000)
$N_e[HCG]$	(40,000; 55,000; 70,000)
Human/Chimp speciation T_{HC}	(3; 4; 5; 6) My
Difference in the two speciation times	(2; 3; 4) My
Divergence with Orangutan d_{HCGO}	18 My;
Mutation rate μ	1.0e-9 mutations per bp per year
Recombination rate ρ	(0.5, 1.0, 1.5) cM/Mb

Table ST4.1: Parameters used for CoalHMM bias correction

1,000,000 sites were simulated in each case, using the GTR+Gamma model estimated on the real data set. The observed bias in the simulations was adjusted using a linear model, which was used to predict the bias in the real data and correct the estimated values.

Demographic model

The input data for CoalHMM is an alignment of genome sequences for each of the four great apes. For any species pair, e.g. human and chimpanzee, at any given position x in the alignment, some component of the total sequence divergence (substitutions per bp) will correspond to mutations occurring on the separate lineages since HC speciation, and the remainder will correspond to ancestral polymorphism, i.e. mutations occurring within the ancestral HC population. We express this as

$$d_{HC}(x) = d_{HCs}(x) + \theta_{HC}(x)$$

where $d_{HCs}(x)$ and $\theta_{HC}(x)$ are the components of sequence divergence associated with speciation and ancestral polymorphism respectively. In essence, CoalHMM uses a speciation model, shown in Fig. SF4.1, to infer this decomposition of sequence divergence in each region of the alignment. The key assumptions of this model are that a) there is no genetic exchange between the species between the time of speciation and the present day, and b) the rate at which individual genetic lineages coalesce in the ancestral population is constant in time.

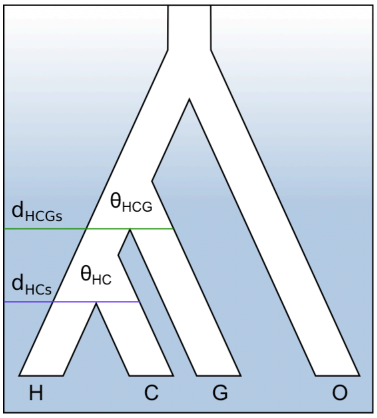


Figure 4.1. Speciation model used by CoalHMM, showing sequence divergences d_{HCs} and d_{HCGs} associated with the HC and HCG speciation times, and ancestral polymorphism parameters θ_{HC} and θ_{HCG}.

Thus in each region CoalHMM infers four demographic parameters: sequence divergences $d_{HCs}(x)$ and $d_{HCGs}(x)$ associated with the HC and HCG speciation times, and ancestral polymorphism parameters $\theta_{HC}(x)$ and $\theta_{HCG}(x)$ of the corresponding ancestral populations, as well as the average recombination rate.

We discuss below the scaling of these sequence divergence estimates using a mutation rate to obtain speciation times in years, and also the consequences of departures from the two key assumptions.

The CoalHMM analysis described above was carried out on both MASKED and FULL datasets, and the final bias-corrected results from both are shown in Table ST4.2, averaged over the autosomes and Chromosome X. As expected, empirical and inferred divergence parameters are slightly lower on the MASKED dataset because it does not contain the fastest evolving regions of the genome. We have used the MASKED values for estimation of speciation times and subsequent analysis, but note that this difference corresponds to an effective average change in mutation rate of approximately 8%.

	FULL		MASKED	
	Autosomes	X	Autosomes	X
Number of sites	1,951,882,490	55,080,167	646,270,079	8,002,847
d_{HCs} (kbp^{-1})	7.96 ± 0.04	6.78 ± 0.28	7.38 ± 0.08	6.44 ± 0.56
d_{HCGs} (kbp^{-1})	12.92 ± 0.03	11.34 ± 0.32	11.9 ± 0.12	10.3 ± 1.1
θ_{HC} (kbp^{-1})	6.97 ± 0.32	3.6 ± 0.32	5.82 ± 0.16	3.12 ± 0.56
θ_{HCG} (kbp^{-1})	3.76 ± 0.08	3.03 ± 0.16	3.09 ± 0.08	2.64 ± 0.24
median ρ (cM/Mb)	1.21 ± 0.80	1.17 ± 0.76	1.91 ± 1.12	1.39 ± 1.00
ILS (%)	30.22 ± 0.41	16.20 ± 2.18	29.65 ± 0.56	12.69 ± 5.31
((H,G),C) / ((C,G),H)	1.06 ± 0.02	1.09 ± 0.07	1.01 ± 0.01	0.87 ± 0.30

Table ST4.2. Parameters inferred by CoalHMM, using both the FULL and MASKED datasets. $d_{s,HC}$ and d_{HCGs} are the components of sequence divergence associated with HC and HCG speciation times (see Section 6 below). θ_{HC} and θ_{HCG} are polymorphism parameters for the ancestral HC and HCG populations. Also shown are the inferred genomic parameters of median recombination rate and total ILS fraction are given along with the ((H,G),C) / ((C,G),H) ILS asymmetry. Errors are s.e.m.

Complete results for all inferred parameters (after bias correction) in 1 Mbp windows of the FULL dataset are provided in Table ST4.3 (Supplementary File).

Variance and statistical noise in CoalHMM's results

The inferred parameters reported in Table ST4.3 are sensitive to mutation rate and ancestral effective population sizes, both of which are expected to vary across the genome from one region to the next. However under CoalHMM's model assumptions, the ratio d_{HCGs}/d_{HCs} should be unaffected by either mutation rate or ancestral effective population size. Table ST4.4 shows the mean μ and standard deviation σ of this ratio on the autosomes and Chromosome X in CoalHMM's output. Since the scale of μ and σ is dependent on demography, we also calculate the coefficient of variation CV = σ/μ.

To establish the level of variance expected from statistical noise alone, we carried out 50 coalescent simulations of 1 Mbp regions generated under CoalHMM's demographic model (using similar parameters to those inferred from the real data) and applied CoalHMM's inference and bias correction processes to the resulting simulated data (output provided in Table ST4.5 (Supplementary File)). In these results we find that the ratio d_{HCGs}/d_{HCs} has a CV which is 67% of that seen in the MASKED autosomal data and nearly 60% of that in the FULL autosomal data. Since this variation is seen in clean simulations under the generative model used for inference, there are additional sources of noise not represented and therefore plausibly accounting for residual variation in the real data. These include that a) the real data contains alignment gaps within each of the 1 Mbp regions used, and in particular after long gaps the Markov chain used in CoalHMM's inference is re-initialised, and b) the sequences present in the real alignment are subject to sequencing and assembly errors, both of which are non-uniform across the genome. The latter in particular is expected to be more notable on X, consistent with the observation that the CV is higher on X in both FULL and MASKED datasets but is substantially reduced in the latter.

An additional source of variance could be that the real data departs from CoalHMM's assumptions in that this ratio is not in fact constant, perhaps due to hybridization or similar complexity in either or both of the HC and HCG speciation events, affecting some parts of the genome more than others. However, we note that this effect if present is not significant on the chromosomal scale: on no chromosome in the MASKED dataset is the mean significantly different from the genome-wide average (data for individual autosomes not shown). We return to this issue below in the context of simulations of migration and ancestral substructure.

	FULL		MASKED		Simulations
	Autosomes	X	Autosomes	X	
μ	1.63	1.70	1.62	1.60	1.57
σ	0.13	0.24	0.11	0.14	0.07
CV	0.082	0.141	0.069	0.090	0.046

Table ST4.4: Mean, standard deviation and coefficient of variation (CV = σ/μ) seen in CoalHMM's estimates of the ratio d_{HCGs}/d_{HCs} on 1 Mbp regions in the FULL and MASKED datasets and in 50 coalescent simulations under CoalHMM's demographic model.

Mutation rates and scaling the results from CoalHMM

The speciation parameters inferred by CoalHMM are expressed in units of sequence divergence (substitutions per bp), so we need a genomic mutation rate (substitutions per year) in order to scale them to units of time. Given such a rate, for example, we can convert the estimates of d_{HCs} and d_{HCGs}, into speciation times T_{HC} and T_{HCG}. However this rate needs to represent an average along the lineages separating the hominines, and there is considerable uncertainty regarding the rate or rates appropriate for hominine evolution. In this section we discuss the factors constraining this choice and the resulting speciation times we obtain.

The mutation rate also varies along the genome from one region to the next. This will manifest as variation in the values of d_{HCs} and d_{HCGs} inferred by CoalHMM, since mutation rate is not explicitly represented in its model. (It will also affect θ_{HC} and θ_{HCG}, but these parameters are also determined by regional variation in ancestral effective population size – see below.) Additional variance in these estimates comes from noise associated with the data processing and coalescent analysis, and some may also be introduced by the presence of ancestral substructure and gene flow – an issue we return to below.

To obtain global speciation time estimates therefore, we use values $d_{HCs}(A)$ and $d_{HCGs}(A)$ averaged over the autosomes, and the mutation rate we scale by is also taken to be an autosome-wide average.

$$T_{HC} = d_{HCs}(A) / 2\mu$$
$$T_{HCG} = d_{HCGs}(A) / 2\mu$$

Chromosome X is excluded as rates are expected to be lower there for several reasons associated with that chromosome's haploidy in males.

Present-day human mutation rate estimates and the generation time

Within the hominid family, mutation rate data is available only for humans, where we have estimates of the rate per generation based on analyses of *de novo* mutations in disease genes and direct sequencing of family trios (Table ST4.6).

	per-generation mean mutation rate (10^{-8} bp^{-1}generation^{-1})	yearly mean mutation rate (10^{-9} bp^{-1}y^{-1})	
		t_{gen} = 30 y	t_{gen} = 25 y
Kondrashov (2002)	1.85 (0.00-3.65)	0.62 (0.00-1.22)	0.74 (0.00-1.46)
Lynch (2010)	1.28 (0.68-1.88)	0.42 (0.23-0.63)	0.51 (0.27-0.75)
Roach et al. (2010)	1.10 (0.68-1.70)	0.37 (0.23-0.57)	0.44 (0.27-0.68)
Awadalla et al. (2010)	1.36 (0.34-2.72)	0.45 (0.11-0.91)	0.54 (0.14-1.09)
1000 Genomes Project (2010), CEU	1.17 (0.94-1.73)	0.39 (0.31-0.57)	0.47 (0.38-0.69)
1000 Genomes Project (2010), YRI	0.97 (0.72-1.44)	0.32 (0.24-0.48)	0.39 (0.29-0.58)

Table ST4.6: Published per-generation mean mutation rate estimates with 95% confidence intervals in studies of *de novo* mutations in modern humans, with corresponding yearly mutation rates assuming 30 year and 25 year mean generation times.

A value of the generation time t_{gen} is necessary to convert rate per generation into a rate per year: $\mu = \mu_{gen} / t_{gen}$. Formally, t_{gen} is the mean separation time between generations averaged over genetic lineages connecting the samples, and recent studies have estimated a generation time of around 30 years in modern humans (Matsumura & Forster, 2008). With this scaling the estimates in Table ST4.6 give yearly mutation rates ranging from 0.32 to 0.62 x 10^{-9} bp^{-1}y^{-1}. However, chimpanzeee and gorilla are observed to have a shorter generation time than modern humans: estimates of around 20 years have been made for both species (e.g. Teleki, Hunt & Pfiffering, 1976; Matsumura & Forster, 2008). It may also be that the 30 year generation time in modern humans is a recent development associated with changes in life history within the last few millenia. Thus we expect a higher mutation rate to apply in the other hominines today and over the three lineages as a whole back to their common ancestor. Based on this, in what follows we consider two values, μ = 0.5 x 10^{-9} bp^{-1}y^{-1} and μ = 0.6 x 10^{-9} bp^{-1}y^{-1} for the mutation rate in hominines, reflecting an expected mean with a generation time of 25 years and a reasonable higher value if the mutation rate is at the upper end of the estimated confidence intervals and/or the mean generation time is lower. We note that these are substantially lower than the commonly quoted value of 1.0 x 10^{-9} bp^{-1}y^{-1}.

Expected time of orangutan divergence requires a higher ancestral mutation rate

Table ST4.7 shows values for T_{HC} and T_{HCG} derived from d_{HCs}(A) and d_{HCGs}(A) using mutation rates of 1.0, 0.6 and 0.5 x 10^{-9} bp^{-1}y^{-1}, under the assumption that a single constant mutation rate has obtained across the whole hominine clade throughout its evolution. The two lower rates are more consistent with several putative hominin fossils (Fig1c) and, as we have seen, are also compatible with estimates of the modern human per-generation mutation rate. However as the table also shows, if we further extend them all the way back to the orangutan divergence, they would imply a very ancient hominid common ancestor (T_{dHO} = 30-36 Mya). Such a timeline is not consistent with the fossil evidence, which indicates a date in the mid-Miocene (12-16 Mya) for the speciation of orangutan from the other hominids, and therefore a common ancestor 20 Mya at the earliest (assuming a comparable ancestral population size to that of the HC ancestor). Earlier times such as those of the hominoid and cercopithecoid speciation events, corresponding to sequence divergences with gibbon or macaque, would be even more at odds with the evidence of the fossil record if scaled with constant rates of 0.6 or 0.5 x 10^{-9} bp^{-1}y^{-1}.

From Fig. 1b, we see that if the average mutation rate on the lineage between human and orangutan were 0.8 x 10^{-9} bp^{-1}y^{-1} this would suffice to bring the inferred HO sequence divergence time down to 19 Mya, with an implied speciation time perhaps 13-16 Mya if the ancestral great ape population size were similar to that of the HC population. Note that this would be consistent with the current mutation rate per year if there had been a substantial slow down in per-year mutation rate since the common ancestor of the great apes.

Mutation rate slowdown in primates

In fact it has been established from comparisons over longer timescales that mutation rates have not been constant during primate evolution, and that primate genome sequence evolution has not followed a molecular clock. In particular the 'hominoid slowdown' – reduced branch lengths within the hominoid (ape) superfamily compared to their closest relatives the cercopithecoids (Old World monkeys) since the common ancestor approximately 30 million years ago – was first proposed by Goodman (1961), and has been confirmed in many studies since then. For example Kim et al. (2006) estimated a relative slowdown of on average 28.4% on the human lineage compared to baboons since the common ancestor, and Steiper & Young (2006) estimated slowdowns ranging from 9% to 21% on hominoid lineages compared to several cercopithecoids. The variation in these and other similar results is partly due to their having been obtained from small amounts of sequence rather than whole genome alignments, but the deviation from a constant rate model has been found to be strongly significant in relative rate tests.

In light of its direct relationship to mutation rate, change in generation time has been proposed as a major factor in the variation of the primate molecular clock, with a mutation rate slowdown corresponding to a generation time increase. Similar life history indicators such as lifespan and age of maturity have also been shown to scale as mass$^{1/4}$ across a wide range of organisms (Gillooly et al. 2005) and in primates in particular (Charnov 2004). We consider here what this implies for changes in generation time within the apes, given the observed changes in body size. Gibbons weigh around 6-12 kg, and Miocene fossil apes are generally estimated to have weighed about 10 kg, whereas present day apes are up to an order of magnitude heavier (from ~40 kg for female chimpanzees to ~180 kg for male gorillas) (Fleagle 1999). This change in body mass by factors of 4-10 would correspond to a generation time increase by factors of 1.4-1.8, and thus yearly mutation rate reduction by 0.55-0.7.

Mutation rate slowdown constrained by molecular clock observations

The actual values of speciation times and ancestral mutation rates under such an explanation depends on the pattern of rate change, which is inaccessible to direct measurement. However we can explore possible models to evaluate whether there is a plausible history consistent with available data.

We noted above that if the average mutation rate μ_{HOav} on the HO lineage were 0.8 x 10^{-9} $\mathrm{bp}^{-1}\mathrm{y}^{-1}$ this would suffice to bring the implied HO speciation time into the mid-Miocene. If the change was linear on both H and O branches from mutation rate μ_{HOav} at the time of divergence T_{dHO} to rates μ_{H} and μ_{O} at t = 0 (the present), such that e.g. on the human branch the rate at time $t < T_{\mathrm{dHO}}$ is

$$\mu(t) = \mu_{\mathrm{H}} + (\mu_{\mathrm{dHO}} - \mu_{\mathrm{H}})t / T_{\mathrm{dHO}}$$

(and similarly on the orangutan branch), then constraining the average rate μ_{HOav} means the ancestral rate is given by

$$\mu_{\mathrm{dHO}} = 2\mu_{\mathrm{HOav}} - (\mu_{\mathrm{H}} + \mu_{\mathrm{O}}) / 2 \qquad (4.1)$$

An additional constraint comes from observations of the sequence divergence of both apes from macaque. Table ST3.3 shows only slightly greater sequence divergence from macaque to orangutan than to the African apes, on the order of 2%, which seems very small and therefore to require a nearly comparable decrease on rate on the pongine branch in parallel to that on the hominine branch. But these numbers are dominated by the long shared divergence from macaque, and in fact allow a substantial difference between hominine and pongine mutation rates. The sequence divergence d_{AM} between macaque and any ape species A, incorporating a linear mutation rate slowdown on the A lineage from ancestral rate μ_{T} to μ_{A} starting T_{dHO} years ago, can be integrated from macaque to A as

$$d_{\mathrm{AM}} = 2T_{\mathrm{M}}\mu_{\mathrm{dHO}} - T_{\mathrm{dHO}}(\mu_{\mathrm{dHO}} - \mu_{\mathrm{A}})/2$$

where T_M is the sequence divergence time of apes and macaques (for which a reasonable estimate is 35 My). From this equation applied to human and orangutan we can derive an expression for the present-day mutation rate in orangutan:

$$\mu_O = \mu_H \frac{d_{HO}}{d_{HM}} + \left(\frac{d_{HO}}{d_{HM}} - 1\right)\left(4\frac{T_M}{T_{dHO}} - 1\right)\mu_{dHO} \qquad (4.2)$$

Thus we can combine the two constraints (equations 4.1 and 4.2) to get:

$$\mu_O = \frac{\mu_H \frac{d_{HO}}{d_{HM}} + \left(\frac{d_{HO}}{d_{HM}} - 1\right)\left(4\frac{T_M}{T_{dHO}} - 1\right)\left(2\mu_{HOav} - \frac{\mu_H}{2}\right)}{1 + \frac{1}{2}\left(\frac{d_{HO}}{d_{HM}} - 1\right)\left(4\frac{T_M}{T_{dHO}} - 1\right)}$$

As above, we take T_{dHO} = 19 Mya and μ_{HOav} = 0.8 x 10^{-9} $bp^{-1}y^{-1}$. The data in Table ST3.3 gives d_{OM} / d_{HM} = 6.35/6.23 = 1.02. If we take μ_H = 0.6 x 10^{-9} $bp^{-1}y^{-1}$, this implies a present-day orangutan mutation rate μ_O = 0.73 x 10^{-9} $bp^{-1}y^{-1}$, and an ancestral rate of μ_{dHO} = 0.94 x 10^{-9} $bp^{-1}y^{-1}$ at the time of HO divergence. This represents a slowdown 37% less strong on the orangutan branch than on the branch leading to human. With μ_H = 0.5 x 10^{-9} $bp^{-1}y^{-1}$ we get μ_O = 0.64 x 10^{-9} $bp^{-1}y^{-1}$ and μ_{dHO} = 1.03 x 10^{-9} $bp^{-1}y^{-1}$, representing a slowdown 25% less strong. Both these would correspond to shorter generation times in orangutans than in hominines, as observed. The corresponding rate reductions of these two scenarios since the Miocene HO ancestor, 0.61 and 0.48 respectively, overlap the lower end of the expected rate reduction calculated above from change in body mass to the power of one quarter.

Thus we conclude that a model of changing mutation rate due to increase in generation time can accommodate both a Mid-Miocene orangutan speciation time and the macaque sequence divergence data, without requiring the degree of parallelism between human and orangutan branches to be overly exact, and with present day mutation rates compatible with the expected range of generation times. An alternate explanation that there is a higher modern mutation rate (i.e. the recent direct measurements of mutation rate are all biased downwards), and less change in mutation rate over time, would bring the HC and HCG speciation events forward in time.

Speciation times under a model of changing mutation rate

We have seen that even with a simple model of linear mutation rate change, given suitable endpoints we can satisfy most or all of the constraints imposed by present day per-generation rate measurements, the date of orangutan speciation, macaque sequence divergences and expected changes in body size. We now consider what this model implies for the dates of HC and HCG speciation.

We derive the following expression by integrating along the H and C lineages from 0 to T_{HC}:

$$T_{HC} = \left(\sqrt{\left(\mu_H + \mu_C\right)^2 + \frac{2d_{HCs}}{T_{dHO}}\left(2\mu_{dHO} - \mu_H - \mu_C\right)} - \left(\mu_H + \mu_C\right)\right)\frac{T_{dHO}}{2\mu_{dHO} - \mu_H - \mu_C}$$

where d_{HCs} is the sequence divergence corresponding to HC speciation (i.e. the number estimated by CoalHMM). A similar equation holds for T_{HCG}, replacing μ_C by μ_G and d_{HCs} by d_{HCGs}. Table ST4.7 shows the resulting estimates of T_{HC} and T_{HCG} under models corresponding to the two different present-day hominine mutation rates we have considered.

mutation rate	T_{HC} (My)	T_{HCG} (My)	T_{dHO} (My)
constant $\mu = 1.0 \times 10^{-9}$ bp^{-1}y^{-1}	3.69 ± 0.04	5.95 ± 0.06	15.5
constant $\mu = 0.6 \times 10^{-9}$ bp^{-1}y^{-1}	6.15 ± 0.06	9.9 ± 0.1	30
constant $\mu = 0.5 \times 10^{-9}$ bp^{-1}y^{-1}	7.38 ± 0.08	11.9 ± 0.1	36
linear change in μ from 0.94 to 0.6 x 10^{-9} bp^{-1}y^{-1} over 19 My	5.54 ± 0.06	8.51 ± 0.08	19
linear change in μ from 1.03 to 0.5 x 10^{-9} bp^{-1}y^{-1} over 19 My	6.13 ± 0.06	9.15 ± 0.08	19

Table ST4.7: Speciation times scaled from the autosomal speciation sequence divergences d_{HCs} = 7.38 ± 0.08 kbp^{-1} and d_{HCGs} = 11.9 ± 0.12 kbp^{-1} inferred by CoalHMM using the MASKED dataset (Table ST4.2). Scaled speciation times are given for different constant values or linear change models of yearly mutation rate. T_{dHO} is the human-orangutan divergence time under the same scaling, expected to be 2-5 My earlier than the HO speciation time depending on the ancestral HO population size.

Considering the putative hominin fossils shown in Fig. 1b, the more recent speciation dates obtained under a slowdown model with a present-day human rate of 0.6 x 10^{-9} bp^{-1}y^{-1} are compatible with hominin status for *Ardipithecus*, but not *Ororrin* or *Sahelanthropus*. The older dates obtained with a present-day rate of 0.5 x 10^{-9} bp^{-1}y^{-1} are compatible with *Ardipithecus* and *Ororrin* as hominins, but still not *Sahelanthropus*. However, it is important to note that while we have discussed a mutation rate slowdown in the context of a linear model, this is only a first-order simplification. For example if most of the change occurred early in hominid evolution, this would result in older speciation time estimates (e.g. close to those obtained with a constant rate of 0.5 x 10^{-9} bp^{-1}y^{-1}), thereby accommodating *Sahelanthropus* as a hominin and possibly also *Chororapithecus* as a gorillin.

Given our uncertainty regarding the level and changeability of ancestral mutation rates, we conclude that the genetic evidence is compatible with a range of HC speciation times from 5.5 to 7 Mya, and HCG speciation times from 8.5 to 12 Mya. Therefore it is consistent with several alternatives concerning the placing of these fossil taxa with respect to human ancestry, and in particular does not rule out hominin status for even the oldest of them. A better knowledge of mutation rates in ancient hominids would be necessary for genetic data to provide strong answers, but it is hard at present to see how such information can be obtained.

Ancestral effective population sizes

Population genetic models are often parameterized by the effective population size N_e, which does not correspond to census population size except in very simple models, but is representative of population scale as experienced by genes, and characterizes many aspects of genomic evolution. In particular it is related to the expected level of polymorphism θ and the mutation rate by $N_e = \theta / 4\mu$, and thus can be calculated from the estimates θ_{HC} and θ_{HCG} for the ancestral HC and HCG populations, given μ. For example, with $\mu = 0.6 \times 10^{-9}$ bp^{-1}y^{-1}, N_{HC} = 122,000 ± 3000 on the autosomes and 65,000 ± 12,000 on X, and N_{HCG} = 65,000 ± 2000 on the autosomes and 55,000 ± 5,000 on X. We consider relative changes in N_e due to selection or demographic changes.

Comparison to the results of Burgess & Yang (2008)

A previous study by Burgess & Yang (2008) analysed a smaller 5-way primate alignment dataset, which included the great apes and macaque. For the most part our results are in good agreement with theirs. In particular their unscaled speciation time estimates are very similar, which is unsurprising since their underlying population genetic model was also similar, as was their method of accounting for variability in mutation rate. They scaled their results by setting T_{HC} = 4 and 6 Mya, corresponding to neutral mutation rates of 0.98e-9 and 0.65e-9 mutations per bp per year respectively. However their estimates of ancestral polymorphism are higher; this may be because their model precludes recombination (and hence variation of the tree topology) within loci.

Effects of more complex demography on CoalHMM's inference

Using coalescent simulations, we investigated the effects of deviation from the demographic model assumed by CoalHMM, considering two plausible contingencies: a period of partial gene flow between the human and chimp populations after their initial separation, and time variation in the ancestral population size.

Gene flow and ancestral substructure

Figures SF4.1-3 show the results of CoalHMM inference on data simulated under conditions of substructure and gene flow (migration). Three different demographic scenarios were used, each comprising three periods after the split from gorilla: first a single ancestral HC population, followed by a period of separation into two populations with genetic exchange between them, followed by two populations with no exchange. The periods of migration were from 2 to 4 My and 2 to 5.5 My, and in each case the levels of migration from H to C simulated were 0.1, 0.01 and 0.001 in units of $4N_e$m, where m is the proportion of the total population migrating each generation (Hein et al. 2005). Other parameters used were rho = 1.5 cM/Mb and a gamma distribution of mutation rate variation (Dutheil et al. 2009).

Fig. SF4.2 shows the effect on inferred speciation sequence divergence. Also shown, for comparison, are the results of simulations with clean splits at 2, 4 and 5.5 My with no subsequent migration. We can see that the inferred speciation divergence in each case corresponds not to the time of the last genetic exchange but to an average over the migration period: for high migration, the inferred value tends to be closer to the end of the migration period (in forward time), while with low migration, it is closer to the time when it starts.

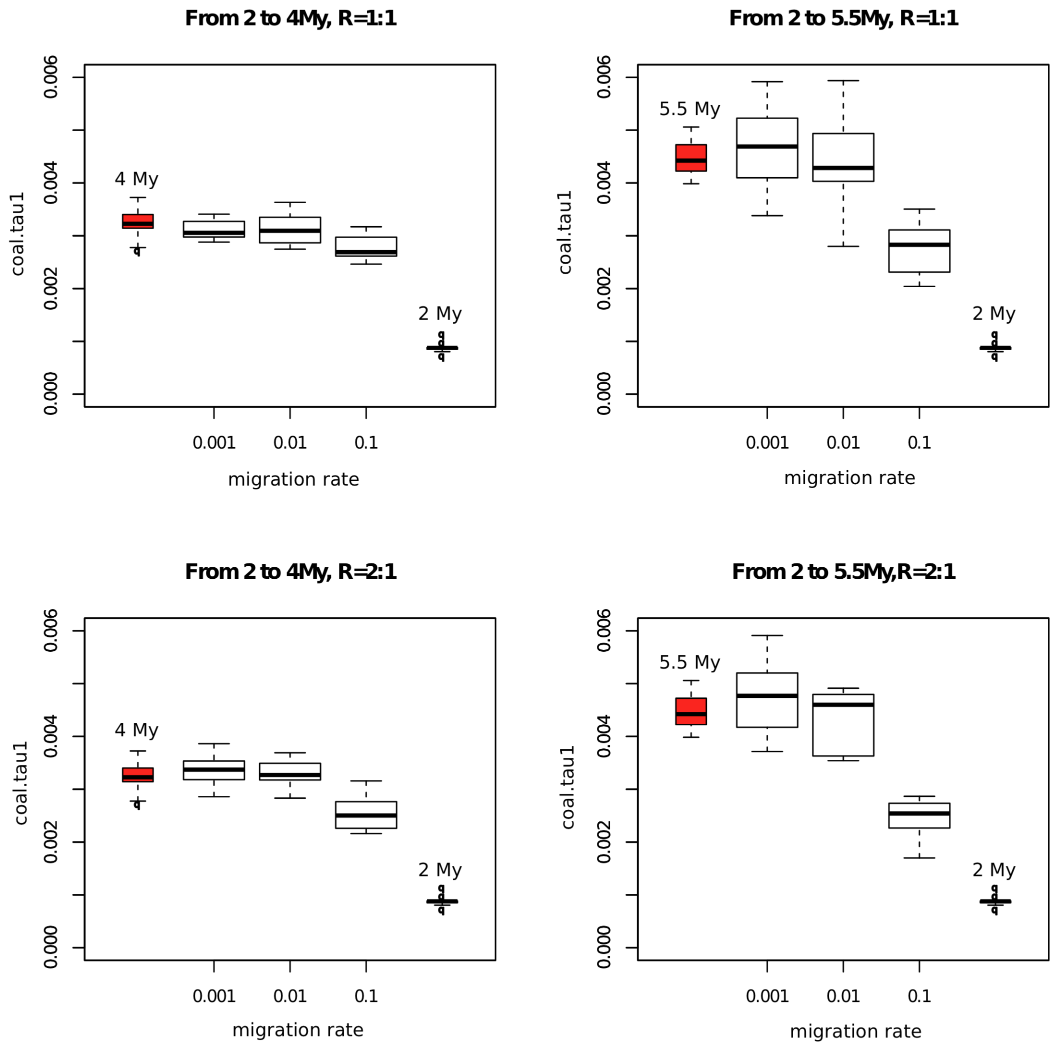


Figure SF4.2. Effects on inferred speciation time of differing scenarios of ancestral substructure around the time of HC speciation. In each plot the x-axis indicates the rate of migration from H to C in units of

$4N_em$, while the y-axis ('coal.tau1') is a non-bias-corrected inferred parameter corresponding to d_{HCs} / 2, in units of mutations per bp. Red boxplots show the results of simulations with clean splits at 2, 4 and 5.5 My, with no subsequent migration. The upper row shows simulations where the migration rates are symmetric between H and C, while the lower row shows simulations where $M_{C\text{->}H} = 2M_{H\text{->}C}$. There are 10 replicates of 1 Mb in each case.

Fig. SF4.3 shows the effect on inferred ancestral polymorphism θ_{HC} (again before bias correction.) In these simulations the value simulated for the population prior to migration was 0.0036; we generally see an increase in inferred θ relative to this. This is in line with expectation, and depends on the interaction between the strength of migration and the time span of the migration period.

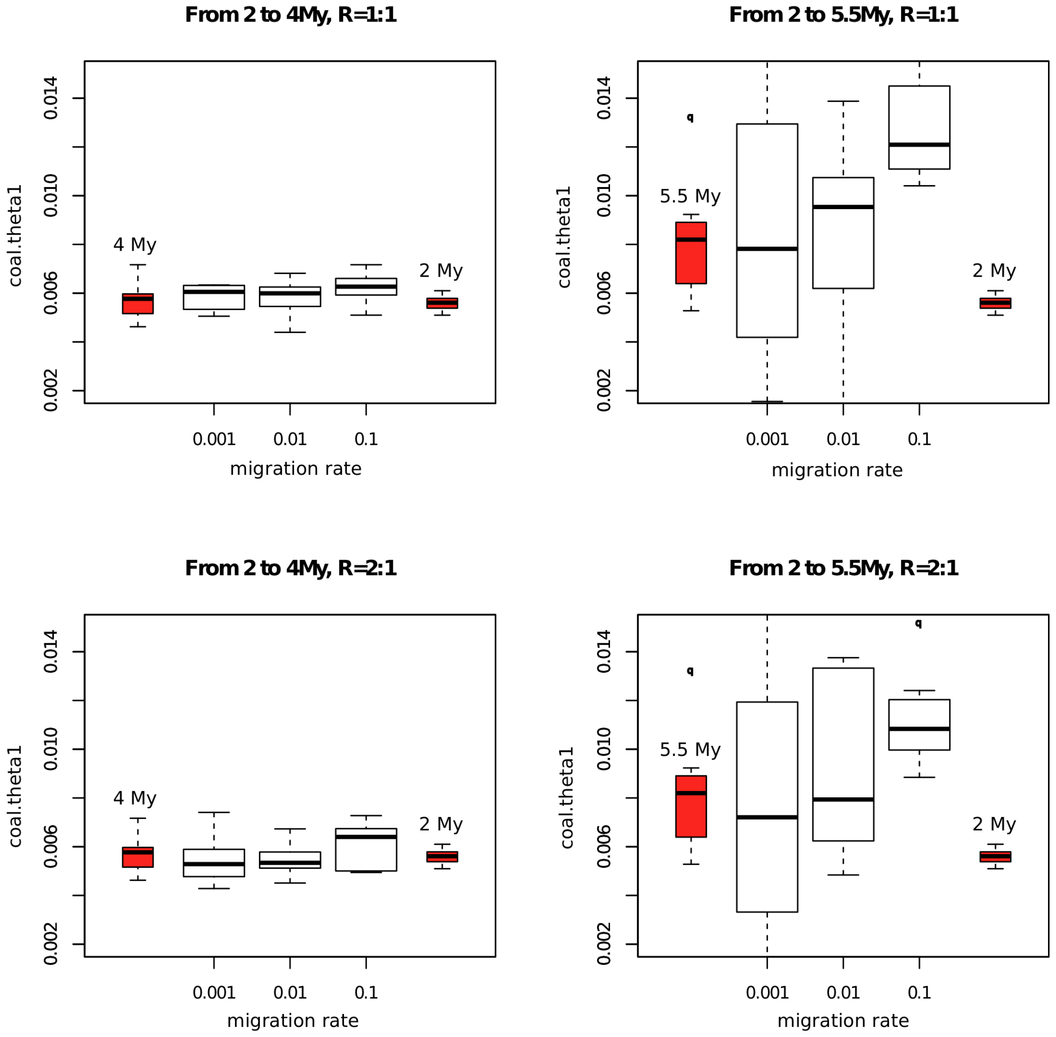


Figure SF4.3. Effects on inferred ancestral polymorphism of differing scenarios of ancestral substructure around the time of HC speciation. As Fig. SF4.2, but here the y-axis ('coal.theta1') is the inferred ancestral polymorphism in units of mutations per bp.

These results show that in conditions of substructure or migration between separated populations, CoalHMM's estimate represents a weighted average, in that when the level of gene flow is high its estimate is close to the end of the period, while when low its estimate is near the start.

We note two points in particular. Firstly, the speciation time T_{HC} is not inferred to be more recent than the time of last genetic exchange between human and chimpanzee, but it may be substantially older, depending on the strength of migration. Secondly, we conclude that ancestral gene flow could result in T_{HC} estimates more recent than those derived from fossil evidence only in a case where strong gene flow caused a recent speciation time estimate but did not significantly affect the development of the derived characteristics observed in the fossil record. It is possible that genetic factors involved in the formation of a population split (perhaps in a parapatric or even sympatric speciation process) could be associated with the morphological characteristics evident in fossil analyses while not inhibiting further gene flow (Coyne & Orr 2004).

More generally, the question arises as to what is meant by speciation time in cases where the separation of two populations was an extended process. One possibility might be to define speciation time as the time of most recent genetic exchange. But this may be impossible to determine in practice, for any particular exchange may well have left no detectable signature either in present day genetic data or in anatomical characteristics. Moreover it could be misleading in cases where there was recent hybridization after a long period of separation. An alternative definition might be the time of initial separation, but this may not be meaningful in cases where the ancestral population contained longstanding substructure or where the separation began slowly and imperceptibly.

For realistic demographies, we argue that it is reasonable to see speciation time - like effective population size - as a model parameter rather than a description of demographic reality. It should perhaps be called effective speciation time. Insofar as it is influenced by phenomena such as migration and the establishment of geographical, behavioural or other reproductive barriers, it represents by a single number the combined effect of diverse evolutionary processes. Therefore we suggest that for genetic data a model-based quantity like the estimate provided by CoalHMM, which is sensitive to gene flow averaged over time, provides a relatively well-defined measure of the time at which two populations separate. Although computationally demanding, it is repeatable and well behaved under plausible demographic scenarios.

Ancestral bottlenecks

We also carried out simulations to investigate the consequences of a bottleneck in the ancestral HC population shortly prior to speciation. This would lead to a large amount of coalescence at the time of the bottleneck, which would break the assumption in CoalHMM that the distribution of coalescent times within the population is exponential. In particular we want to know whether this might reduce the inferred speciation time T_{HC} substantially, perhaps by an amount sufficient to account for the discrepancy in genetic and fossil estimates.

In order to see reasonable ILS values in such a scenario, it is implicit that the HCG and HC speciation times would be closer than without a bottleneck. Our simulations used demographic parameters such that the expected ILS fraction with no bottleneck was 50%, fixing the time of HC speciation at 5.5 Mya, and HCG speciation at 6.6 Mya, with N_{HC} = 90,000. Other parameters were as above. A population bottleneck was introduced lasting from 5.501 Mya to 5.5 Mya, and the bottleneck size was varied: expressed in terms of the ratio of population sizes during and before the bottleneck, it took values 1.0 (no bottleneck), 0.9, 0.75, 0.5, 0.25, 0.1, 0.01, and 0.001. (We did not also vary length as the effects of a bottleneck scale with the product of length and size.) Figures SF4.4-6 show the results.

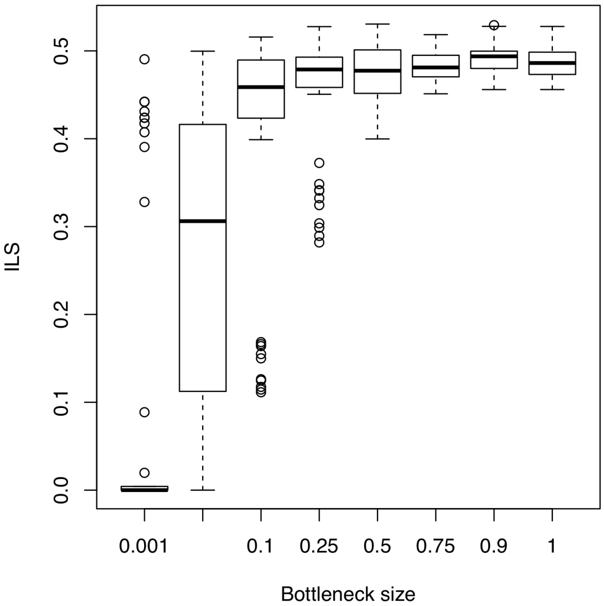


Figure SF4.4. Effect of varying bottleneck size (x axis) on true and inferred ILS fraction.

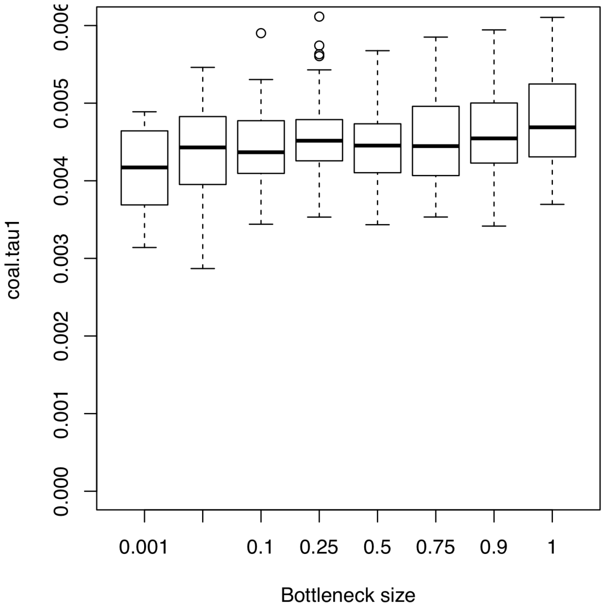


Figure SF4.5. Effect of varying bottleneck size (x axes) on the inferred non-bias-corrected parameter coal.tau1 (as in Fig 4.1).

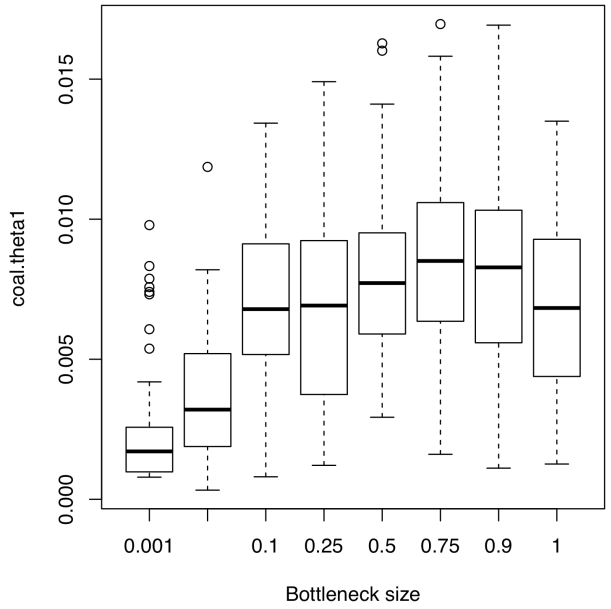


Figure SF4.6. Effect of varying bottleneck size on inferred ancestral polymorphism parameter coal.theta1.

As expected, a bottleneck reduces the amount of ILS in the simulated data (Fig. SF4.4), and a bottleneck size ≤ 0.01 effectively suppresses ILS. We find that CoalHMM's inference of ILS never differs by more than 10% from the true value.

Also as suspected, the inferred speciation divergence is affected by a bottleneck (Fig. SF4.5); however the effect is quite weak, in that even for a bottleneck of size ≤ 0.01 the average reduction is less than 20%. CoalHMM's inference of speciation time is thus robust to variation in the ancestral population size. Instead, as we would hope, the effect of the bottleneck is largely captured within the model by the inferred ancestral polymorphism parameter (Fig. SF4.5).

We conclude that, as with ancestral substructure and migration, variation in the ancestral population size is unlikely to account for the discrepancy between CoalHMM's estimates of great ape speciation times and older estimates based on fossil evidence. Had there been a bottleneck sufficient to bias CoalHMM's estimate by the required amount (e.g. from 6 Mya to 4 Mya), it would also result in negligible ILS, whereas in our data (Table ST4.2) CoalHMM inferred 30% ILS on the autosomes.

5. Incomplete lineage sorting, genome-wide selection, and recombination

Neutral ILS simulations

We can demonstrate that the observed ILS variation (Fig. 2a) is not an artefact of CoalHMM's inference process, and instead reflects genuine variation in the data (e.g. due to ancestral selection), by applying CoalHMM to simulations generated under a neutral coalescent model with uniform ILS probability. We use population parameters matching those inferred by CoalHMM for the great apes, and carry out two sets of 50 simulations of a 1 Mbp region, one with parameters appropriate for the autosomes and one with parameters appropriate for the X chromosome. The results are shown in Fig. SF5.1, along with the corresponding distributions of ILS estimated (also in 1 Mbp regions) from the data.

Leaving aside the issue of model fit between our simulations and the data - the modal values are reasonably well matched, it is clear that the range of ILS values found in the data is much greater, and extends to much lower values, particularly on X. We conclude that the data is therefore inconsistent with a model of genome-wide neutral evolution.

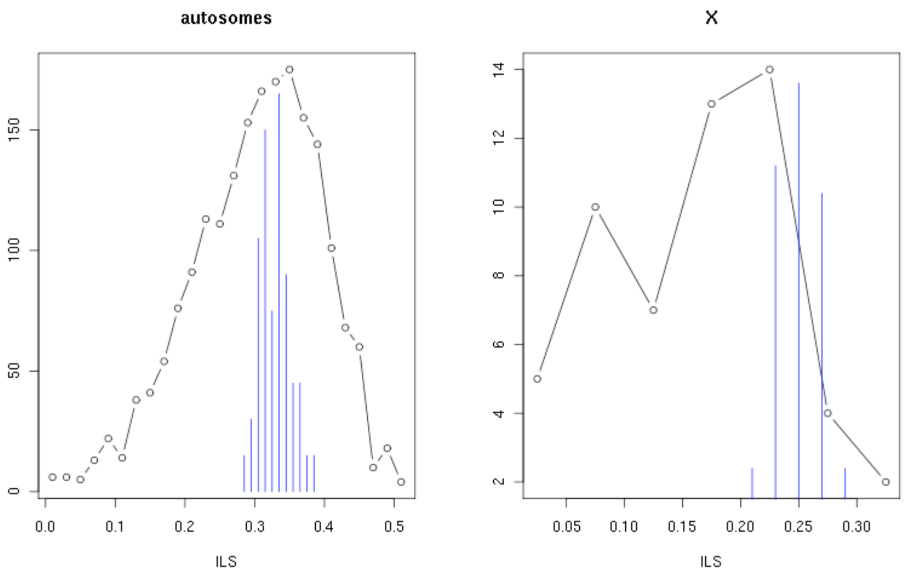


Figure SF5.1: Frequency distributions of ILS estimated in 1 Mbp windows on the autosomes and on X inferred by CoalHMM from great ape data (black) and from simulations of neutral evolution (blue; rescaled so that the modal height matches the black curve).

Correlation of ILS with gene annotations

ILS suppression around coding genes

Taking the MASKED alignment dataset, and labelling alleles in human, chimpanzee and gorilla as 0 if they match orangutan and 1 otherwise, we considered segregating sites with HCG patterns 001, 010, 100, 011, 101 and 110, and counted the number of sites of each pattern in 10 kbp windows across the genome. Within each window we calculated $n_{ILS} = (n_{101} + n_{011}) / h$, where $h = (n_{100} + n_{010} + n_{001} + 2(n_{110} + n_{101} + n_{011})) / 3$. This represents the number of ILS sites normalised by the total tree height - the latter serving as a proxy for the local mutation rate. We then binned these windows according to their distance from the nearest gene transcription start site or stop site (*x*-axis in main text Fig. 2b; negative distances are upstream from a start site). The blue line on Fig. 2b plots the mean value of n_{ILS} averaged over all windows in each distance bin.

Suppression of ILS around genes is not an artefact of assembly errors

A possible concern is that the apparent suppression of ILS sites around genes might be an artefact of assembly errors in the sequences compared, if such errors are substantially more frequent away from genes, and give rise to elevated measures of ILS.

If this were the case, we would expect the human sequence to be the outgroup more often than chimpanzee when counting apparent ILS sites in these regions, since the human assembly is an order of magnitude more accurate than either gorilla or chimpanzee. Fig. SF5.2 plots separately the profiles of site counts with HCGO patterns 0110 (human as outgroup) and 1010 (chimpanzee as outgroup). We see that ILS is reduced at genes to the same extent for both classes of site.

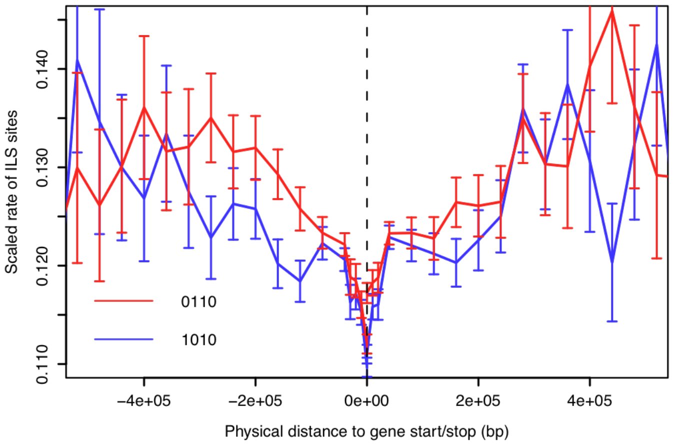


Figure SF5.2: Reduction in ILS around protein coding genes, as Fig 2b, showing separately the profiles for site counts with HCGO patterns 0110 (human as outgroup) in red and 1010 (chimpanzee as outgroup) in blue.

This is consistent with our expectation of error rates in this dataset based on measures of assembly quality. The analysis was carried out on the 'MASKED' dataset prepared for the CoalHMM analysis, which considers only 1:1:1:1 orthologous EPO alignment blocks and excludes RepeatMasked regions, low quality positions in any of the four great ape assemblies and 50 bp windows with many gaps (see Section 4 above). In such regions of the gorilla sequence, the comparison with finished fosmid sequences (Section 3 above) found an average rate of 0.13 single base or short indel errors per kbp. Therefore we would expect 1-2 sequence errors per 10kb interval, compared to the 20-30 ILS sites per interval contributing to the measure that we have plotted. From a comparison of chimpanzee and human sequence divergences from an outgroup, errors in the chimpanzee assembly are also reduced to a comparable level in this dataset (Section 4). Meanwhile errors in orangutan will not significantly affect the detection of ILS sites, since orangutan is used solely to polarize segregating sites in the three hominines. Even the comparison of 1-2 to 20-30 is very conservative over the regions of the genome we have considered, since ILS site patterns (1010 and 0110) require two pairs of matching alleles across the four genomes, which makes it very unlikely for them to be generated by sequencing errors given the error rates indicated above unless the errors are highly correlated between species.

Such correlation could be considered more likely in paralogous regions of the genome (where the duplicated sequences are similar but not identical), which could give rise to spurious ILS sites if the placing of such regions in the chimpanzee assembly differs from that in human, as illustrated in Fig. SF5.3.

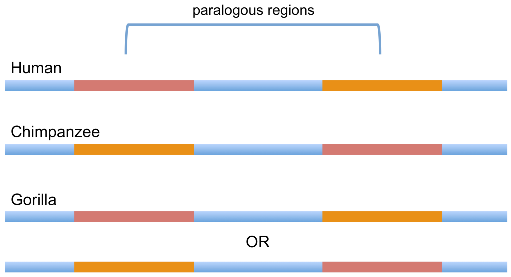


Figure SF5.3: Paralogous genomic regions can give rise to ILS sites if their placing in the chimpanzee assembly differs from that in human (while gorilla matches either of the other two). (Note that not all such cases are necessarily due to assembly errors; some may arise from genuine re-ordering events on either the human or chimpanzee lineages.)

In fact ILS sites generated in this way cannot be responsible for the signal around genes observed in Fig. 2b, as they constitute an insufficient fraction of the total number of ILS sites, independent of distance from genes. This can be seen by considering sites found in regions of the alignment where the human assembly contains a duplication. Table ST5.1 shows counts of ILS sites, segregating sites lying in regions annotated in human as genomic duplications longer than 1 kbp with duplicates of more than 90% similarity (Bailey et al. 2002), and sites in both categories. Firstly, it is clear that at most only 2% of the ILS sites considered are found in these duplicated regions, so that even if all such regions were misassembled as in Fig. SF5.3, so that all ILS sites found therein were spurious, this would still be insufficient to generate the effect shown in Fig. 2b by an order of magnitude. Secondly, although there is a slight enrichment of ILS sites in segmentally duplicated regions, such regions are actually more frequent closer to genes, meaning that the consequences of this enrichment would decrease the effect seen in Figure 2b, rather than contribute to it. A similar correlation of segmental duplication with gene content was reported previously by Bailey et al. (2002) and Zhang et al. (2004).

	Total sites	ILS sites		Sites in human segdup regions		ILS sites in human segdup regions	
distance from nearest gene			% of total		% of total		% of ILS sites
0-50 kbp	306902	8922	0.29	3093	1.01	186	2.08
50-100 kbp	36240	1863	0.51	363	1.00	39	2.09
100-300 kbp	66596	3371	0.51	374	0.56	20	0.59
> 300 kbp	71497	3705	0.52	343	0.48	12	0.32

Table ST5.1: Counts of segregating sites in 1:1:1:1 orthologous blocks of the EPO 6-way alignment categorised by ILS or segmental duplication status. ILS sites are sites with HCGO pattern 1010 or 0110; human segdup regions are regions annotated as duplicated using a whole-genome assembly comparison method (Bailey et al. 2002). Analysis was carried out on 1% of all sites within these blocks genome-wide at which a mutation segregates human, chimpanzee and gorilla.

As further support for the proposition that the ILS sites we have considered are not substantially contaminated by assembly artefacts in duplicated regions, Fig. SF5.4 shows, at ILS sites and at sites supporting the consensus species tree, the distributions of read depth and mean mapping quality of covering reads in the alignment of Illumina data for Kamilah to the gorilla assembly. (Mapping quality is a measure of the confidence with which a read is mapped to a particular genomic location rather than to the most likely alternative, and is therefore expected to be lower for reads mapping to duplicated regions or other areas where the assembly is uncertain.) If ILS sites were significantly enriched for duplicated genomic positions, we would expect the depth distribution at such sites to be wider, and the mapping quality distribution to show greater density at lower qualities. As can be seen in Fig. SF5.4, the distributions are almost indistinguishable.

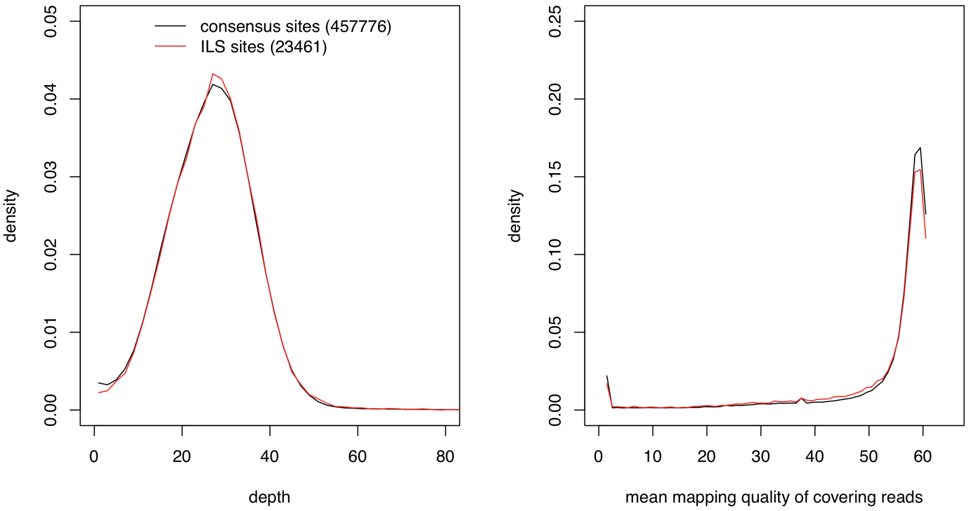


Figure SF5.4: Distribution of depth and mean mapping quality of covering reads in gorilla at ILS sites (red) and sites supporting the consensus tree (black). (Analysis based on 10% of total sites genome-wide.)

Recombination rates inferred by CoalHMM

Lineages separated by a recombination event will eventually re-coalesce. This happens at the same coalescence rate as other coalescence events, so that looking backwards in time, the probability that lineages will have re-coalesced after time t is $1 - \exp(-t/(2*N_e*t_{gen}))$ where t_{gen} is the generation time and N_e is the effective population size. Fig. SF5.4 shows this probability as a function of t for three different values of N_e, with $t_{gen} = 20$ years. For example after one million years only 8% of recombinations avoid re-coalescence with $N_e = 10,000$ whereas with $N_e = 30,000$, 43% remain. After three million years, even with $N_e = 30,000$ only 8% will not have re-coalesced, and since this would be a high estimate for the human and chimpanzee effective population sizes, we therefore expect that almost all recombination events happening on either lineage within the last million years will have re-coalesced before the time of the human-chimpanzee ancestral species. The recombination events we can infer with CoalHMM are therefore expected to have occurred in the ancestral species, or shortly after the speciation event, and not close to the present day.

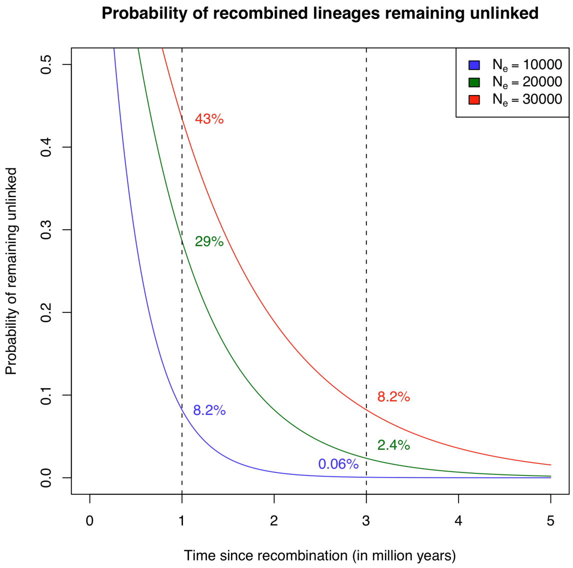


Figure SF5.4: Probability of lineages remaining unlinked.

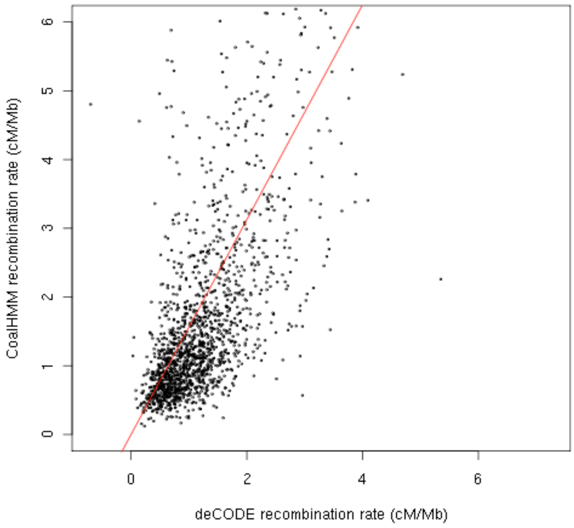


Figure SF5.5: Genome-wide correlation between estimated and present-day human recombination rates. Each point plots the estimated recombination rate in the HCGO alignment and the average rate estimated by REF in a 1 Mbp region. The red line represents a significant correlation (R = 0.49; $p<10^{-13}$).

6. Male mutation bias

At any genetic locus x, the expected amount of ILS between human, chimpanzee and gorilla and sequence divergence between human and chimpanzee can be written as:

$$\mathrm{ILS}(x) = (2/3)\exp(-(T_{HCG} - T_{HC}) / (2N_e(x)t_{gen})) \quad (1)$$

$$d_{HC}(x) = 2\mu(x)(T_{HC} + 2N_e(x)t_{gen}) \quad (2)$$

where $T_{HCG} - T_{HC}$ is the time between the gorilla and chimpanzee speciation events,, t_{gen} is the generation time, $N_e(x)$ is the effective population size of the ancestral HC population at that locus, and $\mu(x)$ the average mutation rate on the lineages connecting H and C. (Both the ancestral effective populations size N_e and the local mutation rate can vary with x.)

The following argument shows how we can use CoalHMM's inference of ancestral N_e to estimate the difference in mutation rate between X and the autosomes – and hence between males and females.

We express (2) as

$d_{HC}(x) = d_{HCs}(x) + \theta_{HC}(x)$

where $d_{HCs}(x) = 2\mu(x)T_{HC}$ and $\theta_{HC}(x) = 4\mu(x)N_e(x)t_{gen}$ are the components of sequence divergence associated with speciation and ancestral coalescence respectively. The CoalHMM model enables us to estimate values for both components of divergence at any position in the five-way alignment. In particular, we can estimate the average autosomal and X-chromosomal speciation divergences $d_{HCs}(A)$ and $d_{HCs}(X)$.

If we assume that the X and autosomal mutation rates differ only on account of their differing modes of inheritance (so that, e.g., while in females the X chromosome experiences the same average mutation rate as an autosome), it is straightforward to show that

$\mu(X) = (2 + \alpha)\mu_f / 3$

$\mu(A) = (1 + \alpha)\mu_f / 2$

where $\alpha = \mu_m/\mu_f$, with μ_m the average mutation rate experienced by a genetic locus in males and μ_f the average rate in females.

Then

$d_{HCs}(X) / d_{HCs}(A) = 2\mu(X)T_{HC} / (2\mu(A)T_{HC}) = \mu(X) / \mu(A) = (4 + 2\alpha) / (3 + 3\alpha)$

Hence

$\alpha = (4 - 3d_{HCs}(X) / d_{HCs}(A)) / (3d_{HCs}(X) / d_{HCs}(A) - 2)$

For the HC comparison in our data (Table ST4.2), using the MASKED dataset we find $d_{HCs}(A)/d_{HCs}(X) = 0.87 \pm 0.08$, corresponding to a male/female mutation rate bias estimate of $\alpha = 2.3 \pm 0.4$. (With the FULL dataset we get $\alpha = 2.6 \pm 0.2$; not significantly different).

For the HCG comparison, $d_{HCGs}(A)/d_{HCGs}(X) = 0.865 \pm 0.09$.

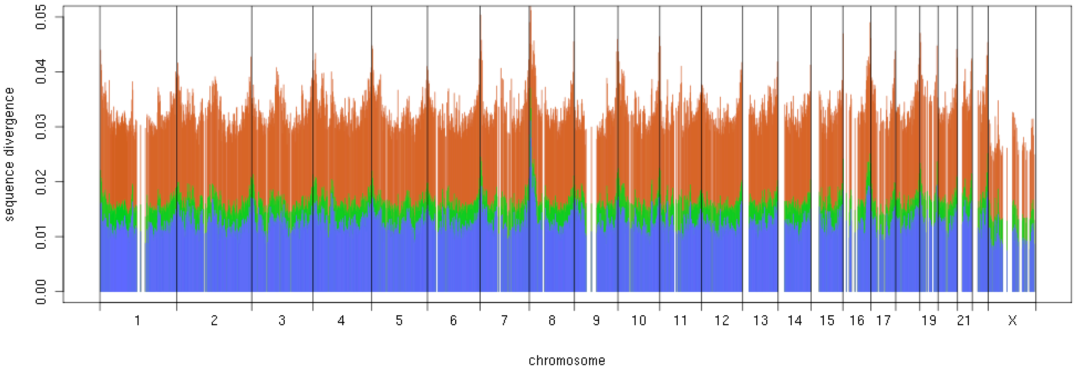


Figure SF6.1: Genome-wide variation in sequence divergence. Each vertical line represents the sequence divergence between human and chimpanzee (blue), human and gorilla (green) and human and orangutan (red), estimated in a 1 Mbp region.

7. Sequence losses and gains within the African apes

We looked for examples of sequence loss or gain within the gorilla, human and chimpanzee lineages by comparing the assemblies, augmented by raw sequence data for each genus. For the latter we used one library of Illumina data at 4x from Kamilah for gorilla, we combined data for a female chimpanzee and a male bonobo to give us 2.5x data representing the *Pan* genus, and we used reads at 7x for NA12878, a human CEU female, from the 1000 Genomes project (1000GenomesProjectConsortium 2010).

Using the alignment program bwa (Li and Durbin 2010) we mapped the data for each genus to three reference assemblies: the reference used for the 1000 Genomes Project (based on ncbi37), the gorilla assembly presented here, and the chimpanzee reference version 2.1. Because repeats in any assembly are particularly sensitive to the assembly method used, we excluded all repetitive material as identified in each reference by RepeatMasker and DustMasker. Hits on the Y chromosome in the human or chimpanzee references were also excluded. (Y is treated separately below.)

In each reference, regions were defined according to their coverage by reads from each of the other two genera. For example, on the gorilla reference, regions were covered either by human reads only, by chimpanzee reads only, or by both. We then excluded all fragments in each reference covered by reads from both other genera. From the remaining material, fragments longer than 100 bp were realigned to each of the three references. If an alignment was found then we identified the material which aligned and re-classified it accordingly. This compensated for the low coverage of the *Pan* data – some material initially classified as not covered by *Pan* was in fact found in the chimpanzee reference at this stage.

The resulting total amounts of non-repeat-masked material found in only one or two of the three species (either in the corresponding reference assemblies or sequence data) are shown in Table ST7.1. This material was then compared to the NCBI RefSeq DNA database using BLAST. Of material found only in the chimpanzee assembly, 43% was not found in any other organism, either indicating a greater incidence of novel material on the chimpanzee lineage or perhaps assembly contamination.

Species in which material is found	Quantity	Distribution of BLAST hit locations (%)			
	(Mbp)	orangutan	Other primate	Non-primate	None
Human only	0.572	80	14	<1	6
Chimpanzee only	0.528	36	15	<1	43*
Gorilla only	1.255	72	14	0	14**
Human & chimpanzee	2.685	80	15	<1	5
Human & gorilla	5.533	81	16	<1	3
Chimpanzee & gorilla	5.562	80	14	<1	6

Table ST7.1: Non-repetitive material found in only one or two of the African great ape genomes, identified in reference assemblies and sequence data, and the distribution of BLAST hit locations for material in each category (excluding hits to the source genome(s)). * Of this (chimpanzee-only material with no hits to other organisms), 38% was from Unplaced reference sequences. ** Of this (gorilla-only material with no hits to other organisms), 47% was from Unplaced reference sequences.

Chromosome	Gene Name	EnsEMBL ID	Protein function
4	C4orf39	ENSG00000250486	uncharacterised
9	LCN10	ENSG00000187922	possible role in male fertility
12	OR6C76	ENSG00000185821	olfactory receptor
19	OR1M1	ENSG00000170929	olfactory receptor
19	OR7G2	ENSG00000170923	olfactory receptor
21	KRTAP13-3	ENSG00000240432	hair keratin associated protein

Table ST7.2: Human Ensembl protein coding genes with exons overlapping material identified as unique to human within hominines.

Chromosome	Gene Name	EnsEMBL ID	supporting evidence
1	-	ENSGGOG00000012254	mouse olfactory receptor protein
2a	-	ENSGGOG00000027212	human interleukin protein
7	-	ENSGGOG00000022569	mouse olfactory receptor protein
11	-	ENSGGOG00000002303	mouse olfactory receptor protein
12	-	ENSGGOG00000025467	human Ret finger protein
unplaced16310	-	ENSGGOG00000023676	human olfactory receptor protein
unplaced2230	-	ENSGGOG00000011802	mouse olfactory receptor protein
unplaced28373	-	ENSGGOG00000011678	mouse olfactory receptor protein
unplaced37942	-	ENSGGOG00000022189	mouse olfactory receptor protein

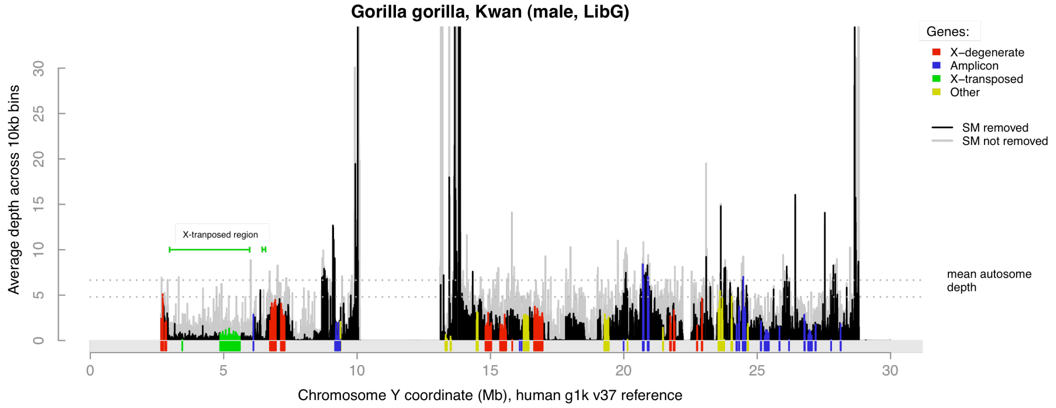
Table ST7.3: Gorilla Ensembl protein coding genes with exons overlapping material identified as unique to gorilla within hominines. No direct functional annotation exists for these genes; the rightmost column gives the functional association of the (human or mouse) protein from which the gorilla gene annotation was derived.

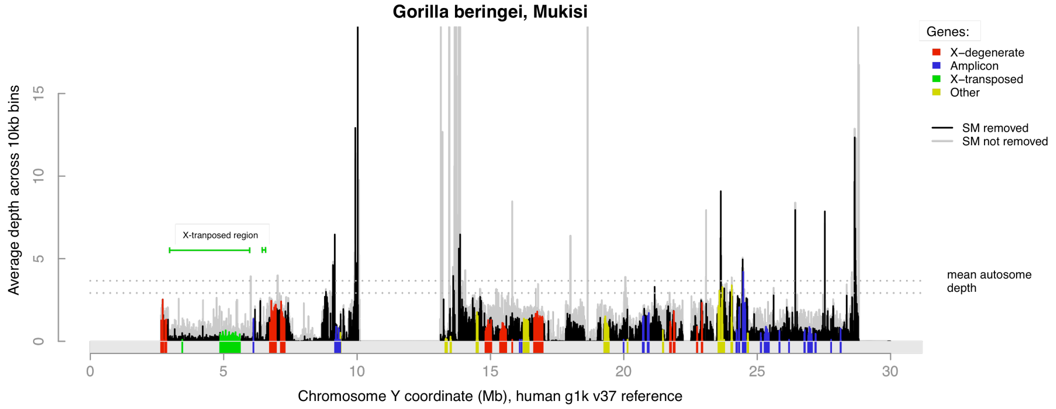
Chromosome	Gene Name	EnsEMBL ID	supporting evidence
12	-	ENSPTRG00000029746	human olfactory receptor protein
16	XM_001170014.1	ENSPTRG00000034534	human uncharacterized protein
17_random	XR_021206.1	ENSPTRG00000034053	human uncharacterized protein
Un	XR_020170.1	ENSPTRG00000031405	human olfactory receptor protein
Un	C11orf89	ENSPTRG00000031380	human uncharacterized protein

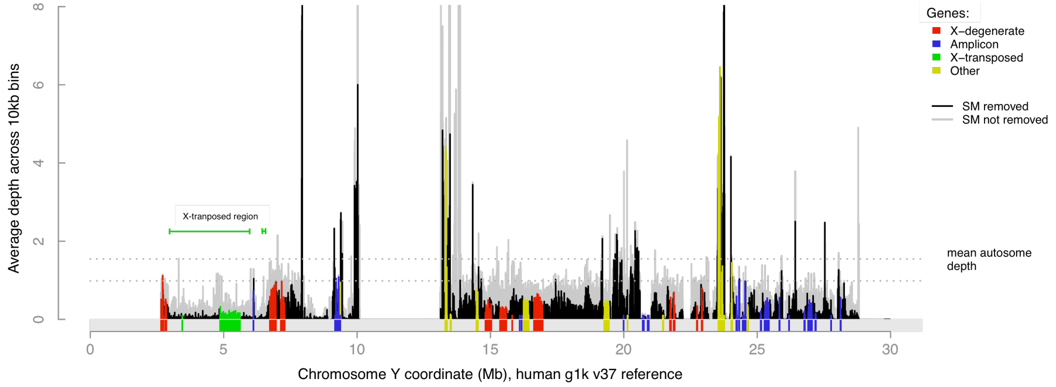
Table ST7.4: Chimpanzee Ensembl protein coding genes with exons overlapping material identified as unique to chimpanzee within hominines. No direct functional annotation exists for these genes; the rightmost column gives the functional association of the (human) protein from which the chimpanzee gene annotation was derived.

Y chromosome

We carried out an analysis of the coverage depth on human Y of reads from the male gorillas Kwanza and Mukisi, the male bonobo and data from the 1000 Genomes Project for a human CEU individual NA10851 with similar coverage to the Kwanza data. The data for each individual was aligned to the whole human reference using bwa, and depth on the Y chromosome calculated in bins of width 10kb. Additionally, we calculated a repeat-mask corrected depth in each bin, excluding repeat-masked bases (and reducing the bin size accordingly when calculating the average depth).


Gorilla gorilla, Kwan (male, LibG)


Gorilla beringei, Mukisi


Pan paniscus, Bonobo1

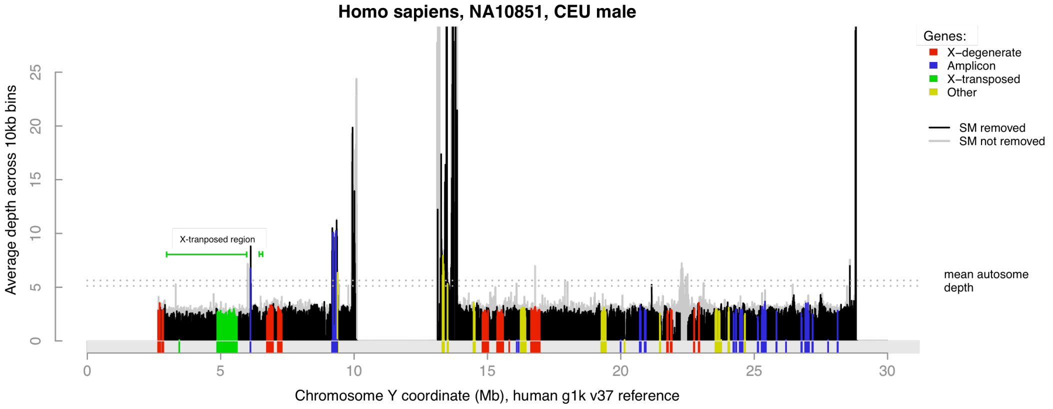
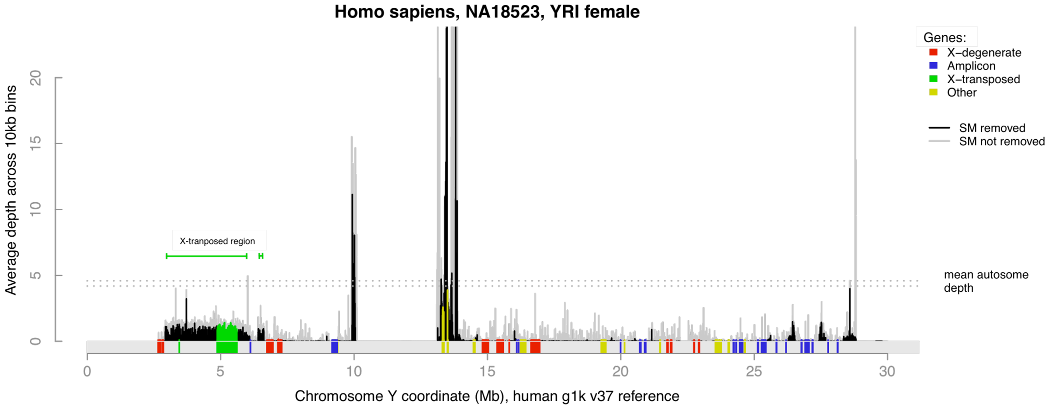


Figure SF7.1: Depth of coverage on human Y of Illumina reads sequenced from two male gorillas, a male bonobo, and a male human, calculated in 10kb bins. In each bin we indicate the average depth (grey) and the average depth corrected for repeat-masked material (black, or colour coded in regions of human protein coding genes in accordance with the sequence classes described in Skaletsky et al. 2003 (Skaletsky, Kuroda-Kawaguchi et al. 2003). Only the first 30Mb of the human Y chromosome is shown (the remainder primarily comprising highly repetitive heterochromatic material). Horizontal dotted lines show the mean (upper line) and repeat mask corrected (lower line) autosomal depth. Note that on the human reference the region around the centromere (between about 10 and 13Mb) is masked out, as is the pseudo-autosomal region (up to about 2.5Mb).

Since there is only a single copy of the Y chromosome in any male individual we expect half the mean autosomal depth on the Y; this is indeed the case in the human male CEU individual (Fig. SF7.1). For comparison, Fig. SF7.2 shows the same analysis on a female human YRI individual NA18523 (again data from the 1000 Genomes Project) and the female Western gorilla Kamilah. After correcting for repeats, very few reads from either individual align to the human Y other than in the X-transposed region (coordinates 3-6 Mbp) and around the centromere. (In particular, for example, there is no coverage over the sex-determining gene SRY - the leftmost red-coloured gene.) Females carry two copies of the region on chromosome X which is homologous to the X-transposed region on the male Y, so we expect reads from these copies to be distributed between both the X and Y regions at half the autosomal depth on each (since the reference contains this material on both X and Y).



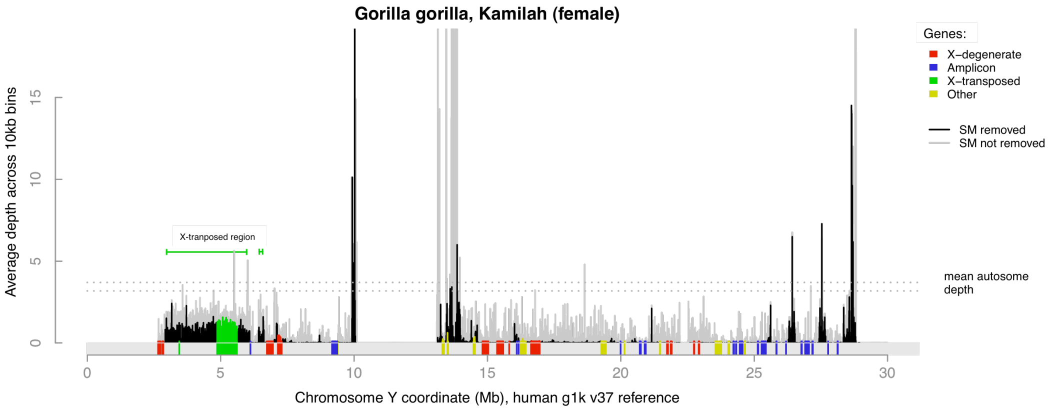


Figure S7.2: Depth of coverage on human Y of Illumina reads sequenced from a female human and Kamilah, a female gorilla. See Fig. SF7.1.

Fig. SL7.1 reveals clear gorilla-human and bonobo-human structural differences. Some regions of the human reference appear to be absent in both gorilla and bonobo; elsewhere the pattern of coverage differs. Additionally, some regions in the ampliconic region seem to be present at greater copy number than in the human reference. On both gorilla and bonobo we observe about one quarter coverage in the X-transposed region, consistent with this region being absent from Y in both species. Since this region is highly homologous with the corresponding region of the X chromosome, a male gorilla or bonobo will have just a single copy of this region (on X), but reads from these regions will align to the human reference in both the X and Y homologous regions, reducing the coverage from 1/2 to 1/4 of the mean autosomal coverage.

For each of the protein coding genes on human Y we determined the average depth of reads mapping to exonic regions within the gene. In Table ST7.5 we list those genes whose exon depth is less than 70% of the expected value.

Individual	X-degenerate	Ampliconic	X-transposed	Other
Kwanza	CYorf15B	VCY, DAZ, TSPY	TGIF2LY, PCDH11Y	AC006156.1, AC134878.1
bonobo1		HSFY, RBMY, DAZ, TSPY	TGIF2LY, PCDH11Y	AC007241.1, AC009235.1

Table ST7.5: Genes on human MSY whose exon coverage in Kwanza and bonobo1 is less than 70% of that expected if present at the same copy number as human. CYorf15B: reported as disrupted in chimpanzee and considered not essential in primates by Hughes et al. (Hughes, Skaletsky et al. 2010)and Goto et al. (Goto, Peng et al. 2009); PCDH11Y, TGIF2LY: Transposed from the X to the human Y after the split from chimpanzees (Page, Harper et al. 1984); TSPY: Multicopy gene which expanded on the human lineage (Xue and Tyler-Smith 2011); DAZ: Four copies in most humans, two in gorillas (Yu, Lin et al. 2008)

8. Protein evolution in the great apes

Identifying primate one-to-one orthologous genes

All ortholog sets, gene trees, and coding alignments were collected from Ensembl Compara release 60 using the Ensembl Perl API (Vilella, Severin et al. 2009; Flicek, Amode et al. 2011). We first identified the set of genes sharing 1-to-1 orthology among all six primates by collecting homology annotations from the Enesmbl Compara database. For each human protein-coding gene, the orthology status for each non-human species was assigned to different categories based on the homologue count and the Ensembl homology annotation: one-to-one (one homologue available and either an *ortholog_one2one* or *apparent_ortholog_one2one* annotation), deleted (no homologue available), duplicated (multiple homologues available), or human duplication (one homologue available but containing an *ortholog_one2many* annotation, indicating that there are multiple human homologs for a single non-human homolog). A table containing the primate orthology status and Ensembl protein IDs for all genes analysed is included in the supplementary file (Table ST8.1). From an initial set of 20,746 human protein-coding genes, this procedure identified 12,652 genes with 6-way 1-to-1 orthology, 4,809 genes with primate deletions, 1,171 genes with primate duplications, 308 genes with human duplications, and 1,806 genes with mixed patterns of duplication and deletion (Fig. SF8.1A).

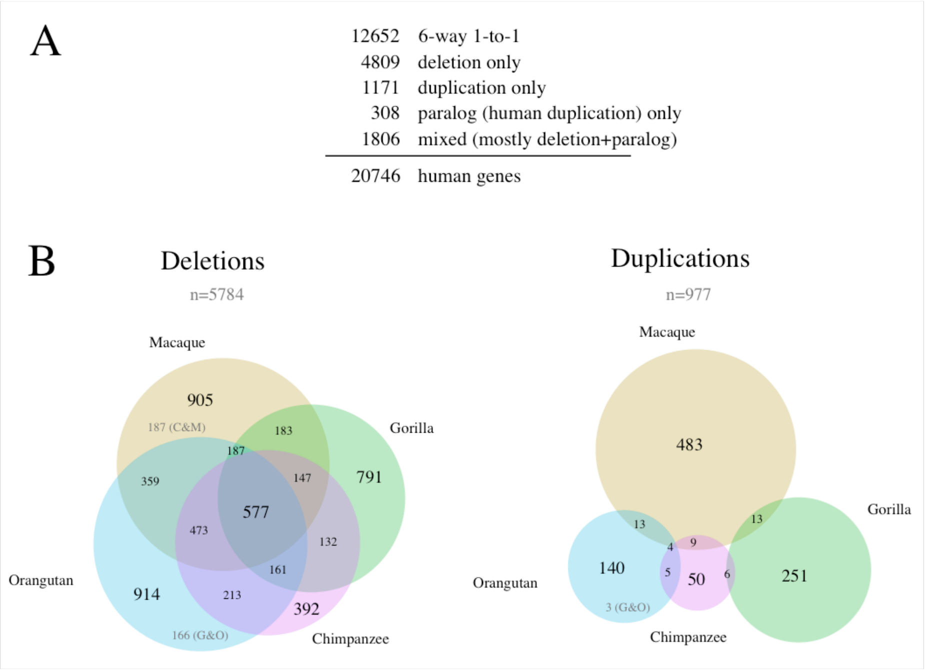


Figure SF8.1: Orthology assignments for human genes based on the Ensembl Compara v60 homology pipeline. **A,** 20,746 human genes were assessed for orthology relationships in six primates and assigned to one of five categories. The 12,652 one-to-one genes were used for subsequent evolutionary analysis. **B,** Venn diagrams showing parallel and lineage-specific deletions in 5,784 genes with at least one primate deletion relative to human and parallel and lineage-specific duplications in 977 genes with at least one primate duplication relative to human. Deletions and Duplications not shown by the approximate Venn diagram topologies are included in grey.

We further analysed the overlap of deleted and duplicated genes in four of the five non-human primates, revealing that roughly half of all genes with at least one Ensembl-annotated primate deletion relative to human had deletions in more than one species (2,782), while only 50 duplications were shared (Fig. SF8.1B). The relative numbers of deletions were largely consistent with each species' phylogenetic distance from human: chimpanzees had 2,279 deletions, gorilla had 2,344, orangutan had 3,050, macaque had 3,015, and marmoset had 3,215. A similar phylogenetic trend was observed for duplication counts, except for a notable excess of gorilla duplications: 273 duplications in gorilla compared to 74 in chimpanzee, 165 in orangutan, 522 in macaque, and 711 in marmoset.

While the Ensembl homology database is likely to be fairly reliable in its classification 1-to-1 orthologous genes, there are many possible sources of error in the identification of species-specific gene duplications and deletions: errors in genome assembly, annotation, alignment, and gene tree inference can all contribute to false positive or false negative duplications or deletions relative to human. The trend towards greater numbers of deletions and duplications with increasing phylogenetic distance to human can be taken as some indication that many of the predicted events are true, but the highly variable nature of primate genome sequence, assembly and annotation quality makes it difficult to overstate the potential error contained within Ensembl's homology identifications. For example, while it would be tempting to speculate that the large number of human genes duplicated in gorilla is due to an exceptionally large amount of duplicated gene-coding material in gorilla, it is likely an artefact attributable to the homology pipeline's gene tree inference step being based on an all-to-all BLAST search and clustering procedure that uses protein sequences derived from each source genome's gene annotations. Since the gene annotations of most primate genomes are guided by homology to the human genome, genomic assembly gaps or breakpoints within regions syntenic to human genes can cause a single non-human gene to be annotated as two separate truncated genes related by an apparent recent duplication event. Given the otherwise highly regular pattern of increased duplication count with increasing phylogenetic distance from human, one may reasonably predict that the elevated number of Ensembl-annotated gorilla duplications will be reduced to a number between that of chimpanzee and orangutan in future releases as the Ensembl gorilla gene annotation set is further refined.

Collecting and filtering six-way codon alignments for one-to-one genes

Codon alignments of all 1-to-1 orthologous genes were collected from Ensembl's 6-way primate EPO genomic alignment set. We extracted and concatenated alignment regions corresponding to the protein-coding portion of each exon comprising the canonical coding transcript of each human gene. The resulting transcript alignment was then flattened to the human reference by removing all columns with insertions in non-human primates or deletions in the human lineage. Since the EPO alignments are generated on the DNA level and our evolutionary analysis was to be performed on the codon level, we cleaned each alignment for codon analysis by masking out any triplets containing stop codons or out-of-frame gaps. Of the original 12,562 1-to-1 genes identified from the Compara homology annotations, 11,538 codon alignments were successfully collected. The drop-off in numbers came from 1,024 genes which were discarded because an entire species was missing from the 6-way EPO alignments; the species most often missing in the alignments were orangutan (520 genes), marmoset (475), and macaque (328).

The low levels of divergence between the primate species being analysed made it extremely important to avoid the inclusion of any incorrectly-aligned material. The expected number of lineage-specific substitutions per gene scales linearly with the terminal branch length, meaning a small number of sequencing, assembly or alignment errors causing apparent non-synonymous mutations along one of the short HCG terminal lineages could easily lead to a false positive inference of accelerated evolution. As such, we applied an aggressive set of filters to each alignment before the evolutionary analysis. First we filtered the chimpanzee, orangutan, macaque and marmoset sequences using PHRED or PHRED-like quality scores downloaded from the UCSC (chimpanzee and macaque) or WUSTL (orangutan and marmoset) websites, replacing any bases with a quality score lower than 30 (corresponding to an expected error rate of greater than 1 in 1000 bases) with Ns. We then applied a filter based on a sliding-window analysis of maximum-likelihood inferred lineage-specific substitutions as follows. The codon alignment was analysed with the *codeml* program of PAML v4.14 using a M0 model in order to infer substitutions in the terminal lineages. Using the branch lengths and substitutions inferred by *codeml*, we analysed the density of codons containing lineage-specific non-synonymous substitutions within 15-codon windows starting at each position of the alignment. The DNA sequence within any window containing more than 10 non-synonymous substitutions per codon per unit of branch length was masked with Ns. A number of heuristic corrections were made to avoid excess stringency or lenience: branch lengths below 0.05 were set to 0.05 in order to avoid too small of a denominator, the threshold was decreased from 10 to 5 for any windows overlapping alignment gaps or ambiguous nucleotides, and codons containing two or three nucleotide substitutions along one branch were counted as two non-synonymous substitutions. This procedure resulted in 72,729 nucleotides being masked from 1,156 alignments, with the following breakdown of numbers of genes in which each species had at least one nucleotide masked: 12 human, 195 chimpanzee, 232 gorilla, 296 orangutan, 271 macaque and 324 marmoset.

The low number of genes from which any human sequence was masked indicates that the filtering was not overly conservative, while the high numbers in non-human primates indicates that those genomes are more likely to contain highly localized, apparently spurious runs of non-synonymous mutations in regions of EPO alignments corresponding to human transcripts. This could either be due to lower-quality genome assembly (causing spurious runs of misplaced sequence to end up in the alignment), errors in the EPO alignments, or pseudogenization events causing a lack of functional constraint and increased numbers of apparent non-synonymous substitutions. To investigate the possibility that the highly-masked genes were enriched for nonfunctional mutations due to pseudogenization, we analyzed the set of highly-masked genes for functional enrichments using the GO enrichment methodology described below, using the 97 genes where the number of masked nucleotides divided by the alignment length was greater than 0.2 as the set of ‘significant’ genes. The results (data not shown) revealed a small number of enriched GO terms indicative of gene functions known to be prone to pseudogenization, including *positive regulation of transferase activity*, *receptor metabolic process,* and *microtubule cytoskeleton organization* (Zhang, Carriero et al. 2004). Those three terms comprised only 27 genes, suggesting that pseudogenes may be present in our codon alignments, but that pseudogenization is not likely a major contributor to the excess of clusters of nonhuman non-synonymous substitutions in our dataset.

One final filter was applied to avoid a potential bias from substitutions in regions of incomplete lineage sorting (ILS) between human, chimpanzee, and gorilla. In the case where a synonymous or non-synonymous substitution occurs along the ancestral branch of a genomic segment subject to ILS (e.g., where gorilla-chimpanzee or gorilla-human share a most recent common ancestry), the assumption of a single phylogenetic tree per gene is violated and PAML cannot properly account for the single substitution event. Instead, two substitutions must be inferred in order to fit the observed site pattern to the phylogenetic tree. The method can choose to infer either two identical substitutions, one along each of the terminal ILS branches sharing the most recent common ancestor, or one substitution along the HCG ancestral branch and a second reversion substitution along the non-ILS terminal branch. We observed that PAML tends to infer the latter sequence of events, likely due to the large branch length between the HCG common ancestor and orangutan making it more likely that a substitution took place on that branch. Regardless, given the high proportion of expected sites under ILS (roughly 20% within gene regions) and the scarcity of multiple independent substitutions at the same site within these closely-related primates, we applied a simple filtering method to mask out codons that were likely the result of a single substitution in an ancestral ILS lineage. Any codon where either gorilla-human or gorilla-chimpanzee shared a codon sequence that was different from both orangutan and the non-ILS species (either human or chimpanzee) was considered likely to contain an ancestral ILS substitution. The human, chimpanzee and gorilla sequences were all masked with Ns at such codons, causing PAML to treat those nucleotides as missing data. This resulted in 7,841 codons being masked from 4,340 genes prior to analysis with PAML. Although the ILS masking was relatively widespread across genes, its effect on the results was conservative with respect to the number of inferred substitutions and likely minimal: the majority of genes (2,605) contained only one masked codon, which would be unlikely to seriously degrade an otherwise strong signal of acceleration or deceleration.

Manual identification of alignment or assembly errors

We identified a few examples of genes with apparent alignment or assembly error that escaped the various filtering steps described above. Candidate genes for erroneous alignment were chosen on the basis of alignment length, branch LRT, and numbers of nonsynonymous and synonymous mutations, and a manual analysis of the alignment was undertaken by visually inspecting the codon alignment, locations of inferred substitutions, and the protein-based alignment of the same gene from v60 of the Ensembl Compara database. Results from these manual analyses are included in Table ST8.2.

- ITPK1, which codes for an inositol triphosphate kinase with 415 amino acids and a primate dN/dS of 0.094, showed 28 non-synonymous and 10 synonymous substitutions in the chimpanzee lineage and a strong signal for chimpanzee acceleration (LRT=26.94). The alignment showed few chimpanzee substitutions in the first half of the protein, but a high density of masked chimpanzee nucleotides and mixed synonymous and non-synonymous chimpanzee substitutions throughout the second half. The substitutions that were not masked by our window-based filter were presumably just below the substitution density threshold. We analyzed the alternative protein-based alignments from Ensembl Compara, which showed a different sequence for the chimpanzee ITPK1 orthologue in the second half of the protein that had far fewer mismatches, suggesting that the EPO genomic alignments from which our data were generated contained a chimpanzee misalignment in the latter half of the protein.
- POLR2A, which codes for the largest subunit of RNA polymerase II with 1971 amino acids and a primate dN/dS of 0.043, showed 22 non-synonymous and 70 synonymous substitutions in the gorilla lineage and a strong signal for gorilla acceleration (LRT=81.2). The gene showed a highly conserved pattern across the bulk of its length, except for the final 250-300 amino acids which contained a highly repetitive sequence and long, dense clusters of substitutions in both gorilla and orangutan. The protein-based alignments showed a much cleaner gorilla sequence but different stretches of substitutions in the chimpanzee sequence, suggesting that this repetitive region is subject to frequent genome assembly and/or alignment error.
- ATN1 codes for a protein of unknown function which, upon the expansion of a trinucleotide repeat within the region, leads to dentatorubral pallidoluysian atrophy, a rare neurodegenerative disorder. With 1191 amino acids and a relatively high primate dN/dS of 0.339, ATN1 showed 105 non-synonymous and 54 synonymous substitutions in gorilla with a strong acceleration signal (LRT=80.24). The gorilla substitutions were evenly interspersed along the length of the gene, but noticeably absent from the first and last 100 amino acids. Gorilla also showed a long 50-amino acid gap before the high-substitution region began. Comparison to Ensembl's protein-based alignments did not show a different gorilla sequence; rather, it contained a chimpanzee sequence with the same pattern of high numbers of substitutions as gorilla. This suggests that there may be a duplicated or pseudogenic version of the gene region in recent primates which contains many substitutions relative to the functional gene, making it likely that the pattern of gorilla substitutions seen in our

ATN1 alignment is not indicative of substitutions acting on a functional ortholog of human ATN1.

- GAS6 codes for a protein thought to be involved in the stimulation of cell proliferation with 722 amino acids, has a primate dN/dS of 0.150, and showed 29 non-synonymous and 8 synonymous gorilla substitutions and a strong signal for acceleration (LRT=20.54). Manual analysis of the alignment shows a block of non-synonymous substitutions in the middle of the gene directly adjacent to a block of sites which were masked by our window-based filter. The Compara protein-based alignments showed a different gorilla sequence in the region of concern, but a qualitatively similar number of differences relative to human, suggesting that either (a) the EPO alignment contained an error causing nonhomologous gorilla material to become aligned to the other primates in the region of concern, or (b) the exon(s) contained within the region of concern are not constitutively expressed in gorilla and have been evolving neutrally. Given the extreme density of substitutions and the apparent lack of homology to human within the region, the alignment error explanation seems more likely.

Two other genes with similarly strong signals of acceleration and numbers of non-synonymous and synonymous substitutions, SUPT16H and POLR1A, were also manually assessed, but no obvious signs of misalignment were found (Table ST8.2). The four genes with clear errors were removed from the lists of genes used for the remainder of the analysis.

Codon model evolutionary analysis

We used the *codeml* program from PAML version 4.4c (Yang 2007) to analyse the evolution of each primate gene using a series of phylogenetic branch and branch-site models. For each model tested, the full length codon alignment was input to *codeml* along with a phylogenetic tree corresponding to the accepted species tree structure (labeled tree, Fig. SF8.2). The 'cleandata' option was set to 0 (e.g., alignment columns containing gaps were included in the analysis), and branch lengths inferred by *codeml* based on the initial M0 model model analysis (assuming one dN/dS ratio across the whole tree and alignment) were used as the initial branch lengths for all other models tested. Substitution events, inferred by *codeml* based on an M0 analysis of the whole alignment after all filtering was applied, were stored for later analysis. Since *codeml* uses unrooted trees and a reversible evolutionary codon model, substitutions along the marmoset branch cannot be distinguished from substitutions along the human-chimpanzee-gorilla-orangutan-macaque ancestral stem and are essentially randomly placed. This ambiguity should not have an impact on the LRT results described here, but will be relevant to the analysis of dN/dS levels using inferred substitution events.

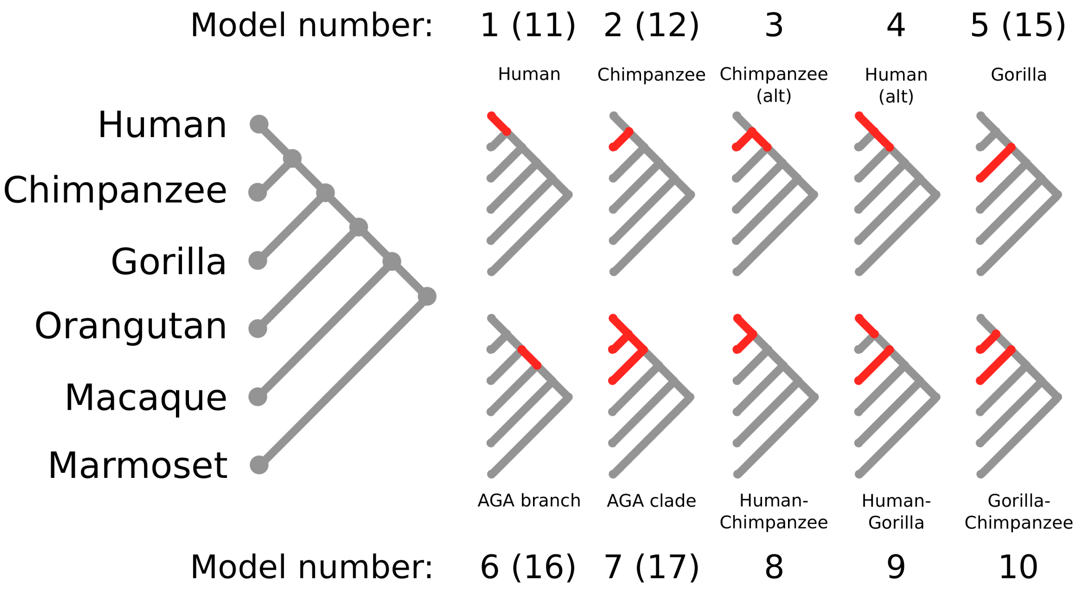


Figure SF8.2: Overview of the branch models used for the evolutionary analysis. Left, the same six-species tree topology was used for all PAML analyses; branch lengths were separately estimated from each alignment. Right, the ten branch models used to test for dN/dS acceleration and deceleration. The same species order as in the tree on left is applied, with foreground branches in red and background branches in grey. Each model was compared to a one-ratio model (not shown) to perform a LRT for increased or decreased dN/dS in the foreground branch(es). Models 3 and 4 were included as alternate models intended to account for the different in branch lengths between the gorilla and chimpanzee/human terminal branches, but models 1 and 2 did not show significant branch length related artefacts and were thus used as the source of human and chimpanzee terminal branch LRTs.

To detect signals of accelerated and decelerated evolution in gorilla and the great apes, we first analysed each alignment with a total of 10 branch models of evolution. Each branch model separated the phylogenetic tree into two categories of branches (foreground and background branches) which were modelled as evolving with separate dN/dS ratios. The branch models used here, numbered 1 through 10 in Fig. SF8.2 (with background branches in grey and foreground branches in red), were designed to allow for elevated (accelerated) or decreased (decelerated) dN/dS ratios in branches of particular interest to the study of African ape evolution. We performed a likelihood ratio test (LRT) for each alignment and model by comparing the likelihood of the alignment under that branch model to the likelihood of the alignment under the simpler M0 model (which assumes a single dN/dS ratio). We call this test the branch-LRT to distinguish it from the branch-site LRT, described below. Using the branch-LRT results, genes were subdivided into those showing evidence for accelerated or decelerated evolution by comparing the estimated foreground and background dN/dS ratio parameters from each optimized branch model; genes where the foreground dN/dS was higher than the background were categorized as accelerations, and genes where the foreground dN/dS was lower than the background were categorized as decelerations. Using this categorization, a signed LRT statistic was constructed for each branch-LRT, where accelerated genes were assigned positive LRT values and decelerated genes were assigned the negative of the LRT value.

A highly positive signed branch-LRT score contains strong evidence for a lineage-specific elevated dN/dS ratio, which could be explained either by positive selection or relaxed constraint. To attempt to distinguish the former from the latter, we used the branch-site test in PAML to identify genes with significant evidence for positive selection acting along a branch or clade. The branch-site test requires a predefined separation of branches into foreground and background categories, similar to the branch test. However, the alternative model of the branch-site test allows a portion of sites in the alignment to evolve with dN/dS > 1 along a specified set of foreground branches, making it specifically tuned towards detecting temporally and spatially localized episodes of positive selection (Nielsen and Yang 1998; Yang and Nielsen 2002; Zhang, Nielsen et al. 2005). The branch-site tests were run using models 11, 12, 15, 16 and 17 (Fig. SF8.2) to allow for comparison to branch model results with the same foreground branches (note that each branch-site model corresponds to the branch model with a number ten below it).

When the null model of evolution in either the branch-LRT or branch-site LRT is true, the LRT statistic - measured as twice the difference in log-likelihood between the branch or branch-site model and the M0 model - should be distributed according to a chi-squared distribution with one degree of freedom (Yang 2007); strictly speaking the branch-site null distribution is a 50:50 mixture of point mass 0 and chi-squared with one degree of freedom, but the more conservative chi-squared distribution is recommended by the author of PAML to guard against violations of model assumptions). P-values for each branch and branch-site LRT were thus calculated by comparing the LRT statistic to a chi-squared distribution with 1 degree of freedom. The method of Benjamini and Hochberg (Benjamin and Hochberg 1995) as implemented in the p.adjust function of the R statistical package (R_Development_Core_Team 2006) was used to adjust each model's set of p-values to correct for multiple testing by controlling the FDR for each test across all 11,538 genes. Signed LRTs and p-values (both raw and adjusted) for the branch tests 1 through 10 (excluding 3 and 4) and branch-site tests 11 through 17 (excluding 13 and 14) are included in Table ST8.5 in columns prefixed with *lrt*, *pval*, and *pval.adj* followed by the branch model. Results from the branch models 3 and 4, designed to detect accelerations along the human and chimpanzee lineages while correcting for the difference in branch length between the human-chimpanzee terminal branches and the gorilla terminal branch, were deemed unnecessary and discarded.

A summary of accelerated and decelerated gene counts and the top ten accelerated genes for each branch model is included in Table ST8.3, and the complete set of LRT results can be found in Table ST8.5. The number of accelerated genes for nearly all branch models was slightly greater than the expected number under the null model, suggesting a mild enrichment for accelerated genes, but the number of strongly accelerated genes varied widely between models. Assuming equivalent amounts of acceleration and deceleration, the expected number of accelerations and decelerations with a branch LRT $p<0.05$ under the null model would be roughly 11,538 * 0.05 / 2 = 288 genes. All models showed an excess of accelerated genes, with between 300 and 873 genes accelerated at the nominal $p < 0.05$ threshold. Decelerations, on the other hand, were either consistent with the null expectation or slightly depleted, with between 151 and 314 decelerations per model. The AGA Stem model showed a notable tendency towards lower dN/dS ratios, with the fewest accelerations (300) and most decelerations (314) out of all models tested. Looking at strongly accelerated or decelerated genes, defined as those with p-values corresponding to an expected FDR < 0.1 after Benjamini-Hochberg multiple testing correction, we observed roughly equivalent numbers in the three terminal lineage models (human / chimpanzee / gorilla), with between 10-19 strong accelerations and between 1-2 strong decelerations. The other models were much more variable, possibly due to differences in power resulting from different foreground branch lengths, with the AGA Clade model showing many strongly-shifted genes (56 strong accelerations and 9 strong decelerations) and the AGA Stem model showing very few (3 strong accelerations and 1 strong deceleration). The Human-Chimpanzee, Gorilla-Human, and Gorilla-Chimpanzee models, designed to detect evidence for parallel accelerations and decelerations, showed roughly twice as many strongly accelerated and decelerated genes as their terminal-branch counterparts (29-45 strong accelerations and 3-6 strong decelerations), as might be expected based on the doubled amount of branch length in the foreground portion of the branch model.

Gene Ontology enrichments

Gene ontology (GO) term annotations for the 'biological process' category were downloaded from v60 of the Ensembl human database (Flicek, Amode et al. 2011) and assigned to the branch model results corresponding to each human gene. Three complementary statistical methods were used to assign p-values for enrichment of GO terms among (a) the most accelerated genes for each branch-LRT performed and (b) genes with evidence of parallel acceleration in a pair of species. For lineage-specific accelerations the 95% chi-squared cutoff value was used to identify accelerated genes, while parallel accelerations for each species pair were identified by genes with a minimum LRT value of 1.5 in both species of interest (more detail on the analysis of parallel accelerations is included in the next section). The first test was a standard one-tailed Fisher's Exact Test (FET) applied to the 2x2 contingency table of significant / non-significant genes which are annotated / not annotated with a given GO term. The second method, implemented in the topGO package for R/Bioconductor, is also based on the FET statistic but additionally compensates for the structure of the GO hierarchy by iterating through the directed acyclic graph and removing nodes from consideration when certain descendant nodes have already shown significant enrichment (Alexa, Rahnenfuhrer et al. 2006). The main effect of the topGO algorithm is to identify and remove semantically repetitive terms from the set of most significantly enriched results by reducing the p-values of terms with more highly-enriched neighbors in the GO hierarchy. The third method accounts for any potential gene length bias in the propensity for a gene to yield a significant LRT results and is implemented in the goseq package for R/Bioconductor (Young et al. 2010). A smoothed probability weighting function (PWF) which predicts the expected proportion of significant accelerations given a gene's length is derived from the genome-wide set of p-values and gene lengths; the PWF is then used to adjust the identification of significantly enriched GO terms to correct for potential over-representation of terms with significantly longer mean gene lengths if longer genes are more likely to yield stronger acceleration LRTs. Although the goseq package was designed primarily for the functional analysis of RNA-seq data (where gene length bias is a widely-acknowledged confounding factor) we observed a small but non-negligible impact of gene length on the likelihood of a gene to contain significant evidence for acceleration and found it useful to run this additional analysis.

The results of the GO enrichment analysis are summarized in Table ST8.4. For each branch model (or pair of species for parallel accelerations) all terms with a FET over-representation p-value below 0.05 and 5 or more significant genes are shown, sorted by their topGO p-value. Any topGO or goseq p-values above 0.05 are colored grey instead of black; terms with non-significant topGO p-values are likely to have a closely-related term with stronger enrichment higher in the list, while terms with non-significant goseq p-values may be enriched by the FET due to a length bias in the detection of accelerated genes.

We note that none of the GO term enrichments for any of the tests remained significant at FDR<0.1 after Benjamini-Hochberg correction for multiple testing. This indicates that none of the branch model accelerations were enriched in GO functional categories at a well-controlled FDR; however, this lack of ontology-wide significance may be due to a variety of factors including the limited power of branch models to detect dN/dS shifts, noise in the GO annotation of genes, or the specific choice of LRT cutoffs in identifying significant accelerations. Our use of a nominal p<0.05 cutoff for enriched GO terms yielded a limited set of enriched terms for each branch model, summarizing the strongest functional associations with moderately to strongly accelerated genes.

Comparison with previous genome-wide evolutionary studies

A number of studies have previously investigated the prevalence and functional associations of positive selection and elevated dN/dS in primates, often using the branch or branch-site models implemented in PAML. Although the studies vary in the exact datasets and analytical methods used, we compared their main results in order to assess the variability of previously published genome-wide results in primates. Table ST8.7 contains a summary of each study including our own analysis for comparison. The proportion of primate accelerations using similar branch model methods ranged from 7.07% to 20.24%; our result, which ranged from 4.64% in chimpanzee to 5.75% in human, is slightly lower than the range of previously-published values, but not strikingly so. For the proportion of genes experiencing positive selection under the branch-site test (or similar tests), our results (ranging from 1.23% in human to 1.66% in gorilla) fall within the lower end of the published range of 0.43% to 8.72%. Most studies do not show a large difference in the proportion of accelerated or positively-selected genes between chimpanzees and humans; our results further confirm this trend and extend the similarity to gorilla.

A wide range of biological functions have been associated with accelerated evolution or positive selection. Terms involving immune functions, olfaction, and amino acid metabolism have commonly been identified in genome-wide scans. Our GO term enrichments based on the branch model results did not recover many terms similar to previous studies; this may be the result of a different sensitivity in the design of our evolutionary models, where gene accelerations in African great ape lineages relative to the primate background rate were detected as opposed to high rates of evolution on their own. For example, immune genes with high dN/dS ratios across all primates were not likely to be identified as accelerated in this study, which could explain the lack of any signal of immune gene enrichment in our results.

Interestingly, we identified enrichments for ‘sensory perception of sound’ in the gorilla lineage and gorilla-human parallel accelerations; although previous studies have detected olfaction and visual perception among functional categories enriched in primate accelerated genes, we believe this is the first genome-wide analysis to provide evidence for an abundance of genes involving sound perception to have been subject to elevated dN/dS ratios in the African great apes. We also found some evidence for enrichment of brain-related terms in human and gorilla, including brain development (gorilla p=0.038) and nervous system development (human p=0.032), although both terms may be subject to a gene length bias and fail to reach p<0.05 using the goseq method for GO term enrichment.

Parallel accelerations in the African great apes

We used the lineage-specific gene acceleration LRT results to evaluate the prevalence and strength of parallel gene accelerations between gorilla and human (GH), gorilla and chimpanzee (GC), and chimpanzee and human (CH) during the time period since the speciation of each pair. We characterised parallel accelerations on three levels, identifying (1) genome-wide levels of shared acceleration, (2) GO terms enriched for shared accelerations, and (3) genes with the strongest evidence of parallel accelerations for each species pair.

Genome-wide rates of shared acceleration

We first identified a suitable statistic by which to measure parallel gene accelerations. Although three of the branch models tested were designed to be sensitive to parallel accelerations in the species pairs of interest (models 8, 9, 10 shown in Figure SF8.2) we found that many of the most enriched genes for these three models were driven by non-synonymous substitutions in primarily one of the two species pairs. For example, the gene with the highest LRT under the gorilla-human model (model 9), SUPT16H, has a LRT score of 51.14. However, it appears that most of this acceleration signal has come from substitutions in the gorilla lineage: looking at the lineage-specific human and gorilla LRTs for the same gene, we find a gorilla LRT of 61.1 and a human LRT of -1.34 (where a negative value indicates an estimated decrease in dN/dS relative to the background branches). Clearly, human and gorilla did not both experience accelerated dN/dS levels in SUPT16H, despite the strong LRT result from the gorilla-human branch model.

To ensure that genes identified as undergoing parallel acceleration showed evidence for having experienced independent dN/dS shifts at a given strength, we used the minimum of both lineages’ independent branch model LRTs (models 1, 2, and 5 in Figure SF8.2 for human, chimpanzee and gorilla, respectively) as the statistic for parallel acceleration in each pair of species. This approach may have slightly reduced power to detect accelerations due to the presence of the paired species in the background model for each lineage (e.g., when detecting gorilla-human parallel accelerations, the human LRT is estimated with gorilla in the background model and the gorilla LRT is estimated with human in the background model). If the gene is truly accelerated in both lineages, then the presence of an accelerated lineage in the background model should reduce the difference in dN/dS levels between the background and foreground models. In this case, however, the influence of the AGA terminal lineages on the estimated dN/dS of the background model was expected to be small compared to that of the orangutan, macaque and marmoset outgroups due to the small fraction of total branch length covered by AGA species.

We used a resampling randomisation strategy to determine whether the number of parallel accelerations at a given LRT cutoff threshold was significantly greater than that expected by chance. For each iteration of the randomisation, we sampled a new set of accelerated genes for each paired species by randomly choosing N genes (where N is the number of observed lineage-specific accelerations for each species at the given cutoff threshold) from among the 11,534 genes in the dataset. We then counted the number of overlapping accelerations at each iteration, and the fraction of iterations which yielded a greater number of overlapping accelerations than the observed number of parallel accelerations was taken as the p-value for the significance of the observed count at the given cutoff threshold. This was repeated for each species pair and for cutoff thresholds ranging from 0 to 5.

We also estimated the magnitude of over- or under-representation of parallel accelerations by calculating the co-occurrence excess (defined as observed score / expected score - 1) of accelerated genes for each pair of species and cutoff thresholds ranging from 0 to 5. We calculated the expected number of parallel accelerations sing the same null expectation as the randomisation test (namely, that each lineage has a characteristic proportion of accelerated genes at a given threshold and that each gene is equally likely to be accelerated). For each species pair and cutoff threshold, one hundred bootstrap replicate datasets were sampled from the 11,534 genes and the co-occurrence excess was calculated for each replicate to generate a distribution of co-occurrence excess values.

The results of the randomisation test and co-occurrence calculations are shown in Figure SF8.3. The co-occurrence excess is plotted as a function of the LRT cutoff threshold for each species pair, with a solid line drawn at the co-occurrence value and a shaded area drawn around the 50% bootstrap confidence interval. The results of the randomisation test are signified by one or two stars drawn adjacent to the co-occurrence value for the given dataset: a single grey star was drawn for datasets with nominal $p < 0.1$, a single black star for $p < 0.05$, and two black stars for $p < 0.01$. For reference, a dotted vertical line is drawn at the LRT cutoff corresponding to a nominal chi-squared $p<0.05$ cutoff for lineage-specific acceleration. At lower (i.e. more lenient) LRT cutoff values more genes in each of the paired species are accelerated, and at higher (i.e. more stringent) LRT cutoff values fewer genes are accelerated; this can be seen in the wider confidence intervals and larger amounts of apparent stochastic noise at higher LRT cutoffs.

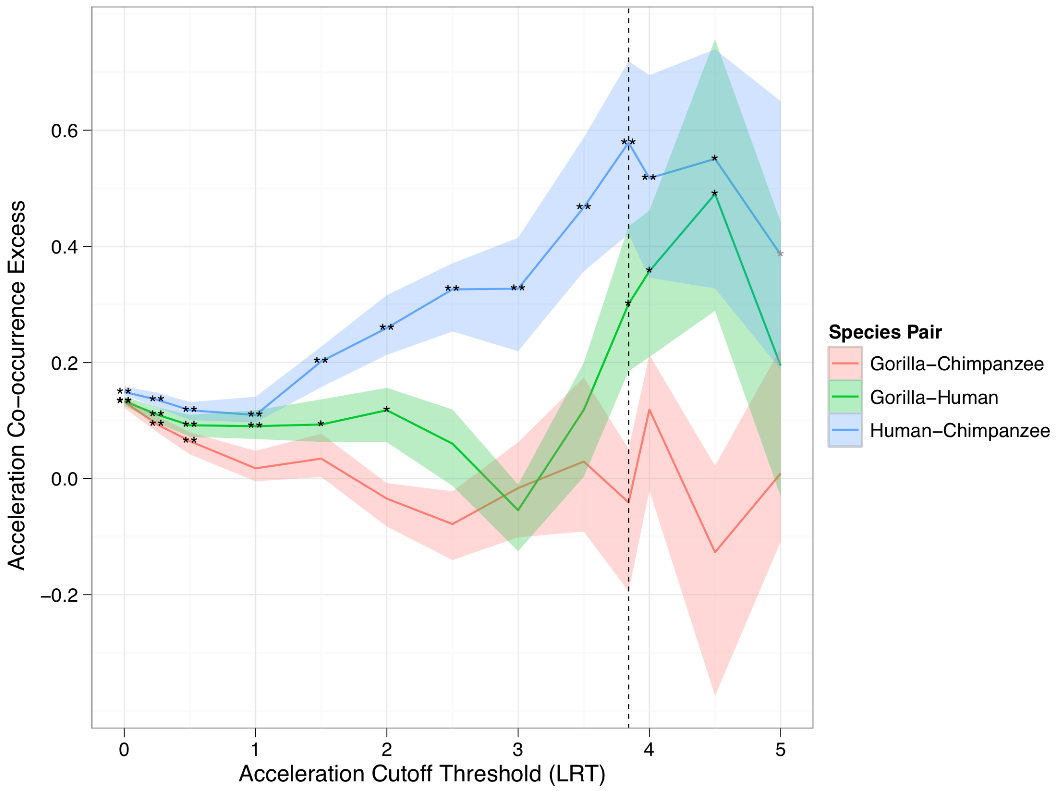


Figure SF8.3. Excess parallel accelerations in pairs of African great apes at various LRT cutoffs. Parallel gene accelerations in three pairs of African great ape lineages (Gorilla-Chimpanzee, red; Gorilla-Human, green; Human-Chimpanzee, blue) were identified using branch model LRT cutoffs from 0 to 5 (x-axis; see text for details on how parallel accelerations were identified), and the co-occurrence excess between lineage-specific and parallel accelerations at each LRT cutoff was calculated (y-axis; the 50% bootstrap confidence interval is shaded). A randomisation procedure was used to identify significant enrichment for parallel accelerations: two black stars indicate $p < 0.01$, one black star indicates $p < 0.05$, and one grey star indicates $p < 0.1$ for the given species pair and LRT cutoff. A vertical dotted line indicates the LRT cutoff corresponding to the 95% chi-squared critical value.

A trend is clear when comparing the co-occurrence excess levels and randomisation test p-values between the GH, GC and HC species pairs: HC shows the largest excess of parallel accelerations, GH shows an intermediate amount of excess, and GC shows the smallest excess. This trend is consistent across a wide range of threshold cutoff values and is supported by both the co-occurrence excess values and the randomisation tests. The HC species pair shows randomisation p-values below 0.01 for all but the two highest (i.e., most stringent) threshold cutoffs and a maximum co-occurrence excess of nearly 60%. The GH species pair shows significantly enriched acceleration overlap at $p < 0.05$ for threshold cutoffs below 2 and at 3.84, 4 and 4.5; the co-occurrence excess was noticeably lower than that of CH, but above zero for all but one threshold cutoff. The GC species pair showed little evidence of genome-wide enrichment for parallel accelerations, with a significant overlap only at weak threshold cutoffs of 0.5 or below and a co-occurrence excess hovering around zero across the range of cutoff thresholds.

That the HC pair shows the largest number of overlapping accelerations is not entirely surprising, as they share the most recent speciation event among the three species pairs and thus presumably share the greatest number of genomic, environmental and behavioral traits that might cause a gene to experience increased non-synonymous mutations in both lineages. More interesting is the difference in overlap levels between the GC and GH species pairs. Whereas the GC pair showed little genome-wide evidence for excess parallel acceleration, the GH pair showed a slight but consistent signal for more parallel accelerations than expected by chance. This could be due either to a greater degree of biological or environmental similarity in GH compared to GC or to some underlying bias in the data, such as differences in population size or genome quality between human and chimpanzee.

GO terms enriched in shared accelerations

We used the same methodology applied in the previous section towards identifying GO terms enriched in gorilla, human and chimpanzee lineage-specific accelerations to identify terms enriched in parallel accelerations for the three species pairs of interest. For parallel accelerations, however, genes where each species had a LRT of at least 1.5 were considered significant; this more lenient threshold was used since few genes were independently accelerated at the 95% chi-squared threshold of 3.84 in both lineages (25 genes for GC, 40 for GH, and 51 for HC). At a LRT threshold of 1.5 the GC pair yielded 206 accelerated genes, GH yielded 238, and HC yielded 286. The GO terms most enriched in parallel accelerations for each species pair are included in Table ST8.4.

Although no species pair yielded GO terms significantly enriched after correction for multiple testing, a number of terms were enriched at a nominal $p < 0.05$ significance using Fisher's exact test. The top three terms for HC parallel accelerations were neuropeptide signaling pathway, regulation of DNA-dependent transcription and microtubule-based movement; for GC parallel accelerations, protein autophosphorylation, inner ear development and positive regulation of transcription factor activity; and for GH parallel accelerations, Wnt receptor signaling pathway, sensory perception of sound and skeletal system morphogenesis.

Interestingly, the term sensory perception of sound was enriched at $p<0.05$ (Fisher p-value) in all three species pairs, though only three genes (LOXHD1, CDH23, GPR98) were significant at LRT > 1.5 in all three pairs -- in other words, all other significant genes were unique to each species pair. The sound perception genes uniquely significant in the GH pair were EYA1, USH1C, MYO3A and SLC1AC; for the GC pair OTOF and FZD4; and for the HC pair DIAPH1, MYCBPAP and DFNB31. This suggests that the tendency of genes involved in sound perception to experience mild to moderately elevated dN/dS levels is relatively widespread in all three African great apes, with variation in the specific genes having undergone acceleration in each species or species pair. It is also worth noting that the enrichment for this term among parallel accelerated genes was strongest in the GH species pair, where it retained $p < 0.05$ in both the topGO and goseq tests. In both the GC and HC species pairs the term had $p > 0.05$ for those tests, indicating that human and gorilla share a slightly stronger signal for parallel accelerated evolution in hearing genes than do the other possible pairs of African great apes.

Top parallel accelerated genes

Using the same statistic for parallel acceleration as used for the genome-wide analysis and GO enrichments, we identified the top parallel accelerated genes for each species pair and for all three species. These genes are listed in Table ST8.3B below the top lineage-specific accelerated genes. Among the top genes accelerated across all three lineages are LOXHD1, a gene comprised of PLAT domains recently shown to be necessary for auditory hair cell function (Edvarson et al. 2011), ITIH3, a plasma serine protease inhibitor potentially involved in prevention of tumor metastasis and associated with risk of myocardial infarction (Ebana et al. 2007), and PARP3, a member of the ADP ribosyl transferase family which has recently been characterized as playing a role in telomeric stability, response to DNA damage, and neural crest development (Rouleau et al. 2011, Boehler et al. 2011). The molecular and medical evidence for the functional activity of these genes suggests that the elevated dN/dS levels across all three African great apes are not well explained by a substantial loss of functional constraint; other possible causes might be relaxed evolutionary constraint due to a lower recent effective population size or positive selection due to functional adaptation, or some combination of the two. Continued study of the evolution of these candidate parallel accelerated genes, particularly with attention paid to the recent population structure of the African great ape species, will further elucidate the shared and unique behavioral and environmental features of our species and our closest evolutionary neighbours.

Assessing levels of incomplete lineage sorting within and near genes

The close proximity of the H-C and HC-G speciation events makes ILS a prominent feature of the genomic relationship between human, chimpanzee and gorilla. Regions showing ILS should not be subject to any particular functional constraint, but the effective population size in the great ape ancestor is expected to affect the prevalence of ILS, causing less ILS in regions of lower ancestral effective population size (Hobolth, Dutheil et al. 2011). Strong purifying selection on genes tends to reduce the effective population size within and nearby exonic regions, so we sought to characterize the effect of the strength of purifying selection, as measured by the primate dN/dS ratio, on the prevalence of ILS within and around protein-coding regions.

We first looked at patterns of ILS density within protein-coding genes. We separated the 11,538 one-to-one genes into five equally-sized dN/dS bins, using the dN/dS ratios estimated from PAML's M0 model on the codon alignments. Before performing the ILS masking step, each alignment site of each gene was assigned a substitution pattern according to the similarity of the human, chimpanzee and gorilla nucleotides to the orangutan nucleotide, using a '1' for similarity and a '0' for dissimilarity and concatenating the values in the order of human, chimpanzee and gorilla. Thus, a site where human and gorilla match the orangutan sequence but chimpanzee is different would be denoted '010'. We counted and stored the total number of sites with each pattern for each gene, skipping all sites with a gap or ambiguous nucleotide in any species. The most important patterns allowing for discrimination of ILS versus non-ILS sites were '110' (likely due to a substitution in the human-chimp ancestor), '101' (likely due to a substitution in a human-gorilla ILS ancestor), and '011' (likely due to a substitution in a chimp-gorilla ILS ancestor). Due to the scarcity of substitutions along the short ancestral branch, we identified genes with strong evidence of ILS by the presence of more ILS patterns than non-ILS patterns -- specifically, when n_101 + n_011 was greater than n_110. A total of 1,974 genes (17.1%) showed evidence of ILS at this threshold, which corresponded well with the results of the CoalHMM analysis observing ILS at 20% of gene-coding sites. To generate the plot shown in Fig. SF8.4A, we calculated the fraction of ILS genes within each dN/dS bin as well as a 95% confidence interval based on 1000 non-parametric bootstrap replicates, plotting the fraction of ILS genes with the black line and showing the 95% confidence interval with each rectangle. This analysis revealed a subtle but clear correlation between ILS density and level of purifying constraint, with the ILS density ranging from 16% in genes with the lowest dN/dS to 19% in genes with the highest dN/dS. This is consistent with the expectation that ILS is depleted from regions subject to persistent purifying selection and shows that the overall level of constraint on a gene has a detectable, if limited, impact on the effective population size of the coding region.

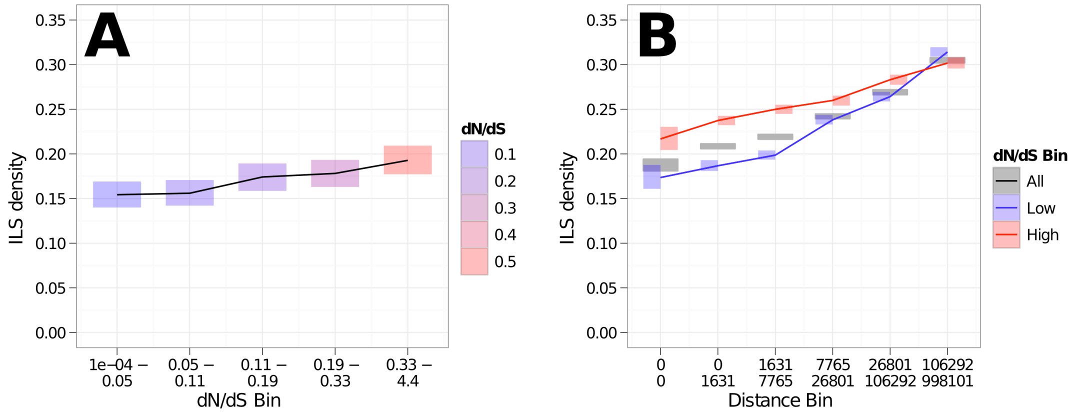


Figure SF8.4. ILS patterns within and surrounding genes. **A,** Genes were separated into five equally-sized dN/dS bins and scored for sites showing patterns of ILS. Genes with lower dN/dS values show significantly lower amounts of ILS, showing the influence of long-term purifying selection on the ancestral effective population size. **B,** Overlapping 10 kbp windows of primate genomic alignments were scored for sites showing patterns of ILS. Windows near protein-coding exons show significantly lower amounts of ILS, with the strongest effect in windows adjacent to genes under strong purifying selection.

We used a similar approach to investigate the effect of gene dN/dS levels on ILS density in regions surrounding protein-coding genes. Counts of ILS and non-ILS patterns were collected for slices of Ensembl's 6-way primate EPO alignments corresponding to 10kb windows of the human genome spaced at 3.3kb intervals, yielding 470,344 genomic regions. Each region was further annotated with the distance from the region's centre point to the nearest protein-coding exon (assigning a value of zero if the region's centre point was within an exon) and the identify of the gene containing the nearest exon. Windows were separated into equally-sized bins according to their distance to the nearest exon (with a separate bin specifically added for windows with a distance of zero) and the plot shown in Fig. SF8.4B was generated by performing the same ILS density calculation for windows within each bin as described in the previous paragraph and plotting the 95% confidence intervals for ILS density in each bin with grey bars. A total of 24.8% of windows across all distance bins were classified as having evidence of ILS. The same calculations were then repeated for windows whose nearest gene was in the lowest 25% quantile (dN/dS < 0.05, blue bars and line) and in the highest 25% quantile (dN/dS > 0.24, red bars and line). The effect of background selection on gene-flanking regions of the genome can be clearly seen in the reduced ILS density in windows near to exons; furthermore, this effect scales with the strength of purifying selection on a given gene and can be observed in windows further than 100kb from the nearest exon.

Global dN/dS ratios in primates and ancestral lineages

The gorilla genome also provided an opportunity to examine global trends in the evolutionary dynamics of the African great apes and their ancestral populations. Specifically, we used the genome-wide set of highly filtered codon alignments to examine estimated dN/dS levels across the six-primate phylogenetic tree.

Theory predicts that larger populations should exhibit, on average, lower dN/dS values due to increased efficacy of purifying selection (Ellegren 2009), and results from several genome-wide analyses have consistently confirmed this trend (ChimpanzeeSequencingandAnalysisConsortium 2005; Lindblad-Toh, Wade et al. 2005; Gibbs, Rogers et al. 2007; Kosiol, Vinar et al. 2008; Warren, Hillier et al. 2008). Within primates, there exists some disagreement regarding the relative dN/dS of human and chimpanzee: Gibbs et al. (Gibbs, Rogers et al. 2007) found a mean dN/dS of 0.175 for chimpanzee and 0.169 for humans, while Kosiol et al. (Kosiol, Vinar et al. 2008) found a mean dN/dS of 0.245 for chimpanzee and 0.249 for humans.

We used the *codeml* program to estimate genome-wide dN/dS levels for each branch in the six-species primate phylogeny from the 11,538 one-to-one alignments. Two genome-wide coding alignments were created: an unfiltered alignment was created by concatenating the alignment of each gene after sequences were filtered for sequence quality but before the window-based filter for clustered non-synonymous substitutions and the filter for ILS-patterned substitutions were applied, and a filtered alignment was created by concatenating the final alignment of each gene that was used for the acceleration analysis. Both alignments were 7.263 million amino acids in length. Before being input to *codeml*, columns containing a gap character or 'N' in any species were removed. As expected, more columns were removed from the filtered alignment due to additional 'N's from the window-based masking procedure; the final unfiltered alignment contained 5.909 million amino acids and the final filtered alignment contained 5.895 mi llion amino acids. A single dN/dS value was first estimated for each alignment using *codeml*'s M0 model, yielding 0.220 for the unfiltered and 0.218 for the filtered alignment. Branch-specific dN/dS values (with standard errors) were then estimated from each alignment using *codeml* with the free-ratios model (parameter 'model=1'). Because *codeml* uses reversible models of evolution and requires unrooted trees as input, any estimates of dN/dS on the outermost branch (connecting the H/C/G/O/M ancestor to marmoset) would be unreliable; thus, we considered marmoset as an outgroup in this analysis.

The resulting genome-wide estimates of dS (the number of synonymous substitutions per synonymous site) and dN/dS for each branch are given in Table ST8.6 and shown in Fig. SF8.5. We find that human has a slightly but significantly higher overall dN/dS than both chimpanzee and gorilla (dN/dS=0.256, 0.249, 0.239, respectively) and that orangutan, macaque, and the ancestral lineages all have lower overall dN/dS values than the terminal branches of the African great apes (ranging from 0.195 for macaque to 0.211 for the human-chimpanzee ancestor). We note that our results are in strong agreement with equivalent estimates from Kosiol et al. (Kosiol, Vinar et al. 2008), who found overall dN/dS values of 0.249, 0.245, and 0.191 for human, chimpanzee, and macaque, respectively.

Branch label	PAML dS	filtered dN/dS	(SE)	unfiltered dN/dS	(SE)	% dN/dS change (unfiltered vs. filtered)
Human	0.0057	0.256	0.00248	0.257	0.00248	0.39%
Chimpanzee	0.0055	0.249	0.00247	0.259	0.0024	4.02%
H/C ancestor	0.0019	0.211	0.00373	0.212	0.00306	0.47%
Gorilla	0.0077	0.239	0.00203	0.249	0.00207	4.18%
H/C/G ancestor	0.0087	0.201	0.00178	0.202	0.00177	0.50%
Orangutan	0.0160	0.213	0.00132	0.218	0.00134	2.35%
H/C/G/O ancestor	0.0131	0.203	0.00147	0.203	0.00147	0.00%
Macaque	0.0324	0.195	0.000887	0.197	0.000888	1.03%
All*	NA	0.218	0.00038	0.220	0.00038	0.92%

Table ST8.6: Global dN/dS values in five primate species. Estimates were obtained from concatenated alignments of one-to-one orthologs from six primate species before (unfiltered) and after (filtered) filters were applied for clustered non-synonymous substitutions and patterns of ILS substitutions. The most distant aligned species, marmoset, was used as an outgroup. *Numbers for the All branch are based on the one-ratio (M0) codon model.

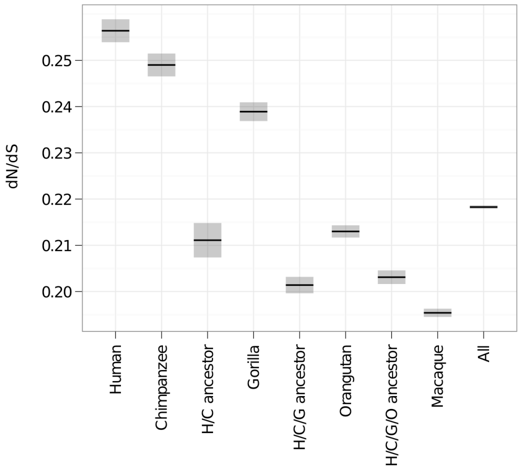


Figure SF8.5: Global dN/dS values in five primate species. Genome-wide dN/dS values were obtained by analyzing a concatenation of filtered one-to-one orthologous alignments from six primate species with codeml using the free-ratios model. Each estimated dN/dS value is plotted as a horizontal line surrounded by a grey box corresponding to the *codeml* error of the estimate. Ancestral lineages are labelled with the first characters of their descendant species.

A comparison of global dN/dS values from the unfiltered versus the filtered alignments in Table ST8.6 reveals that the unfiltered alignments generally gave higher dN/dS values, though the magnitude of change varied dramatically between lineages. The human lineage and all of the ancestral lineages showed less than a 1% decrease in dN/dS, while chimpanzee, gorilla, orangutan, and macaque all showed greater than 1% decrease in dN/dS. Since the window-based masking procedure was only applied to substitutions in the terminal branches, one would not expect the ancestral lineages to experience a significant change in dN/dS; this was confirmed by the small magnitude of change in the ancestral branches. Furthermore, the smaller magnitude of change in human (0.39%) compared to the other extant primate genomes (ranging from 1.03% to 4.18%) indicates that the filtering procedure resulted in the removal of far more non-synonymous substitutions from non-human sequences than from human. Since the masking procedure was applied equivalently to all terminal lineages in a branch-length weighted fashion, the most likely explanation for this discrepancy is that, relative to other primate genomes, the human genome contains fewer sequencing or assembly errors that led to apparent clusters of non-synonymous mutations in Ensembl’s genomic EPO alignments. Interestingly, the magnitude of the dN/dS shift for non-human primate genomes appears to correlate negatively with the length of the terminal branch, with a ~4% shift for gorilla and chimpanzee, a ~2% shift for orangutan, and a ~1% shift for macaque. This observation is consistent with the expected trend if each genome contained a similar number of erroneous bases, because the larger number of true substitutions in species with longer terminal branch lengths would “dilute out” the signal of elevated dN/dS resulting from falsely-aligned bases. In sum, the comparison between filtered and unfiltered dN/dS levels strongly validates the use of the window-based substitution filter, showing that the application of the filter hardly affected the dN/dS of the finished-quality human genome but resulted in branch-length dependent decreases in dN/dS in nonhuman primate genomes with lower-quality assemblies. Notably, chimpanzee showed a marginally higher dN/dS than human in the unfiltered alignment and a significantly lower dN/dS than human in the filtered alignment.

An additional consideration in the interpretation of global dN/dS values is the fact that genes are evolutionarily heterogeneous entities, with each gene composed of sites evolving under different amounts of purifying and positive selection due to varying functional and biological constraints. Since our estimates of dN/dS were based on a codon model with a constant dN/dS value across all sites, they represent the most likely dN/dS value if all sites were evolving under the same selective pressure. In some sense they can be considered an ‘average’ dN/dS value resulting from the combination of sites under purifying, nearly-neutral, and positive selection into a single alignment. Two problems arise as a result of this averaging across genes and sites: first, the comparison of absolute dN/dS values from different studies is difficult, since the specific genes chosen for analysis can have a significant impact on the overall results (Ellegren 2009). Second, the inclusion of sites with very different selective pressures might decrease the signal-to-noise ratio of the data with respect to evaluation of the impact of effective population size of different lineages on the efficacy of purifying selection. Although population theory predicts (and empirical studies have verified) that purifying selection in genes should be more efficient with larger effective population sizes (resulting in lower overall dN/dS values), a subset of protein-coding sites may evolve neutrally or under positive selection. Neutrally-evolving sites in genes should show no relationship with effective population size and positive selection should show the opposite effect, with higher dN/dS values in larger populations due to increased efficacy of positive selection (Ellegren 2009). Although our data show a strong component of the expected signal from purifying selection, the concatenation of heterogeneously evolving sites into the same alignment could have effected the strength of such a signal.

To address this issue, we sorted protein-coding sites into bins based on their sitewise selective pressures in mammals and separately analyzed the set of sites from each bin. Sitewise selection pressures were collected for each site in all of 11,538 genes under investigation by applying the Sitewise Likelihood Ratio (SLR) method (Massingham and Goldman 2005) to coding alignments of the set of Eutherian mammal orthologous genes from the Ensembl Compara database. We used the sitewise likelihood ratio test (LRT) statistic calculated by SLR to sort all sites by their level of evidence for non-neutral selection, with strongly purifying sites (e.g. with low dN/dS) receiving low values, neutral sites receiving intermediate values, and positively-selected sites (with high dN/dS) receiving high values. Sites were split into five bins corresponding to the following quartile ranges of the LRT statistic: 0 to 0.05, 0.05 to 0.33, 0.33 to 0.67, 0.67 to 0.98, and 0.98 to 1.0. These ranges were chosen so that three bins of roughly equal sizes covered the bulk of sites, while two bins focused on the 5% most strongly purifying sites and the 2% of sites with strongest evidence for positive selection. All sites from each bin were concatenated into one alignment and analyzed with *codeml* as described above.

Results from the lineage-specific analysis of sites binned by selective pressure are shown in Fig. SF8.6, with the human dN/dS for each bin plotted on the x-axis (note the log-scale) and the dN/dS of each species relative to human plotted on the y-axis. The 4 lowest dN/dS bins showed the same general trends in dN/dS levels as the combined analysis, with human, chimpanzee, and gorilla showing the highest dN/dS values, followed by orangutan, and finally a cluster of macaque and the ancestral lineages with the lowest dN/dS values. The H/C/G/O ancestral lineage appears to have evolved with a slightly lower dN/dS than the other ancestral lineages, though this difference is only apparent in the 2nd and 3rd bins. Moving from bins with lower human dN/dS to bins with higher human dN/dS, we note a trend across all lineages towards increased dN/dS values relative to human. This is consistent with the expected decrease in the differential effects of population size in sites subject to weaker purifying selection. In the 4th bin, where human has a dN/dS ratio of ~0.55, the cluster of ancestral lineages shows nearly the same dN/dS as orangutan, and the human, chimpanzee and gorilla values are only marginally distinct from each other.

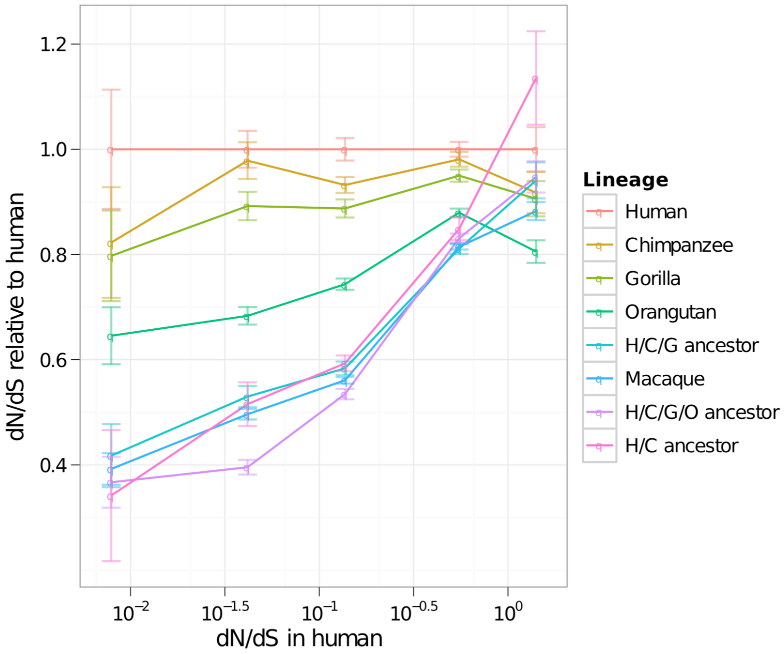


Figure SF8.6: Global dN/dS values in five bins according to selection pressure. Sites were sorted by mammalian selection pressure, assigned to one of five bins (more details in text), and concatenated and analyzed with codeml using the free-ratios model. Estimates are plotted with the human dN/dS for each bin on the x-axis and the lineage-specific dN/dS relative to human on the y-axis. Lines connect estimates from each species, and standard errors of the estimates are shown with error bars. Note the log scale on the x-axis.

The pattern of dN/dS levels in the highest bin -- representing the top 2% of sites ordered by the LRT statistic, and thus the 2% of sites for which evidence of site-specific positive selection across Eutherian mammals is greatest -- is distinct from the other 4 bins and warrants special mention. The human dN/dS in this bin is ~1.4, confirming that this subset of sites does indeed contain a number of sites subject to positive selection. Most of the lineages show a continuation in this bin of the general trend of increasingly similar dN/dS estimates between lineages, with values clustered between ~88%-95% of the human dN/dS. However, the human-chimpanzee ancestral lineage and the orangutan branch show strikingly increased (human-chimpanzee) and decreased (orangutan) dN/dS values in the highest bin. Whereas the human-chipmanzee branch was at ~85% of the human value in the 4th bin, its value was ~110% that of human in the highest bin; on the other hand, orangutan went from ~90% in the 4th bin to ~80% in the highest bin. These two lineages' strong divergence from the trends observed in the other lineages and other bins suggests that some effect other than a difference in effective population sizes has caused an increase in human-chimpanzee, and a decrease in orangutan, of the prevalence of non-synonymous substitutions in sites with evidence for positive selection across Eutherian mammals. One approach to investigating this artefact more deeply would be to identify the subset(s) of genes which contribute most strongly to these lineages' deviations from the trend.

Human-gorilla protein variants

Variants creating premature stop codons in gorilla

To evaluate gene loss in gorilla, we used Ensembl's 6-way primate EPO genomic alignments (Build 37, Ensembl 58) to identify mutations that create premature stop codons (stop gained consequence) in gorilla compared with human, also requiring there to be no paralogue. After manual reassessment we excluded calls made in regions of local misalignment (with ≥ 4 substitutions and/or indels in the 10 bp sequence surrounding the variant) and genes that had 2 stop gained mutations less than 10 bp apart and those with ≥ 3 stop gained mutations, as well as cases where there were multiple adjacent nucleotide substitutions in gorilla leading to a missense rather than nonsense mutation. Table ST8.8 lists the final set of 90 mutations present in 84 gorilla genes.

Variants associated with disease in humans

Table ST8.9 lists cases in which a protein variant thought to cause inherited disease in humans is the only version found in all three gorillas for which we have genome-wide sequence data. Additional capillary (Sanger) sequencing was undertaken to check protein coding variants identified in gorilla which are thought to cause inherited disease in humans. In addition to Kamilah, Kwanza and Mukisi, several further western and eastern gorilla samples were used for this.

Chr:Position gorGor3	Gene	*Hs* Ref	HDA[1]	*Gg* Ref	WESTERN						EASTERN	
					Kamilah (Reference)		Kwanza		Murphy (Gg013)		Mukisi	
					Illumina	Sanger	Illumina	Sanger	Illumina	Sanger	Illumina	Sanger
Chr5:39797090	*GRN* [2]	C [2]	T [2]	A [2, 3]	AA	A	AA	A	No Call	A	AA	A
Chr5:44556816	*TCAP* [2]	G [2]	A [2]	T [2]	TT	T	TT	T	No Call	T	TT	T
Chr16:210823	*HBA1* [4]	C	A	A	AA	A	AA	A	No Call	A	No Call	A

Table ST8.10. Sequencing of gorilla variants associated with inherited disease in humans. Notes: [1]HDA= Human Disease Allele; [2]GRN and TCAP are coded on opposite strands in humans and gorillas so the gorilla allele is the same as the disease allele; [3]The GRN "A" allele is also seen in 4 additional Western (Effie, Floquet, Fubu, Ruby) and one additional Eastern (Victoria) gorilla samples; [4]annotated as "Novel protein coding" in Ensembl.

9. Fosmid analysis

Sequencing and gene content.

We sequenced and finished 13 fosmids covering two groups of genes on chromosomes 22 and 16 (Fig. SF9.1)

The first group aligns to chromosome 22 and contains the following protein coding genes: GH1, GH2, Four CSH genes (CSH1 & 2), CD79B, APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3G, APOBEC3H. The growth hormones (CSH & GH) exist as serial duplications. Compared to the orthologous region in the human genome there is an additional copy of the CSH gene. One CSH gene is truncated by a gap in the fosmid sequence. The APOBEC genes have conserved synteny when compared to the human orthologous region. APOBECF is missing but is predicted to sit within a gap in the fosmid and has been found in an isolated scaffold in ensembl. The position of a number of pseudogenes is also conserved in the human genome.

The second group maps to chromosome 16 and contains the following protein coding genes: TNP2, PRM3, PRM2, PRM1, APOL3, APOL4, MYH16. There are no gorilla specific duplications in this cluster. MYH16, which is a pseudogene in humans, appears to have an ORF but is truncated at the 5' end (c.650bp) at the edge of the fosmid.

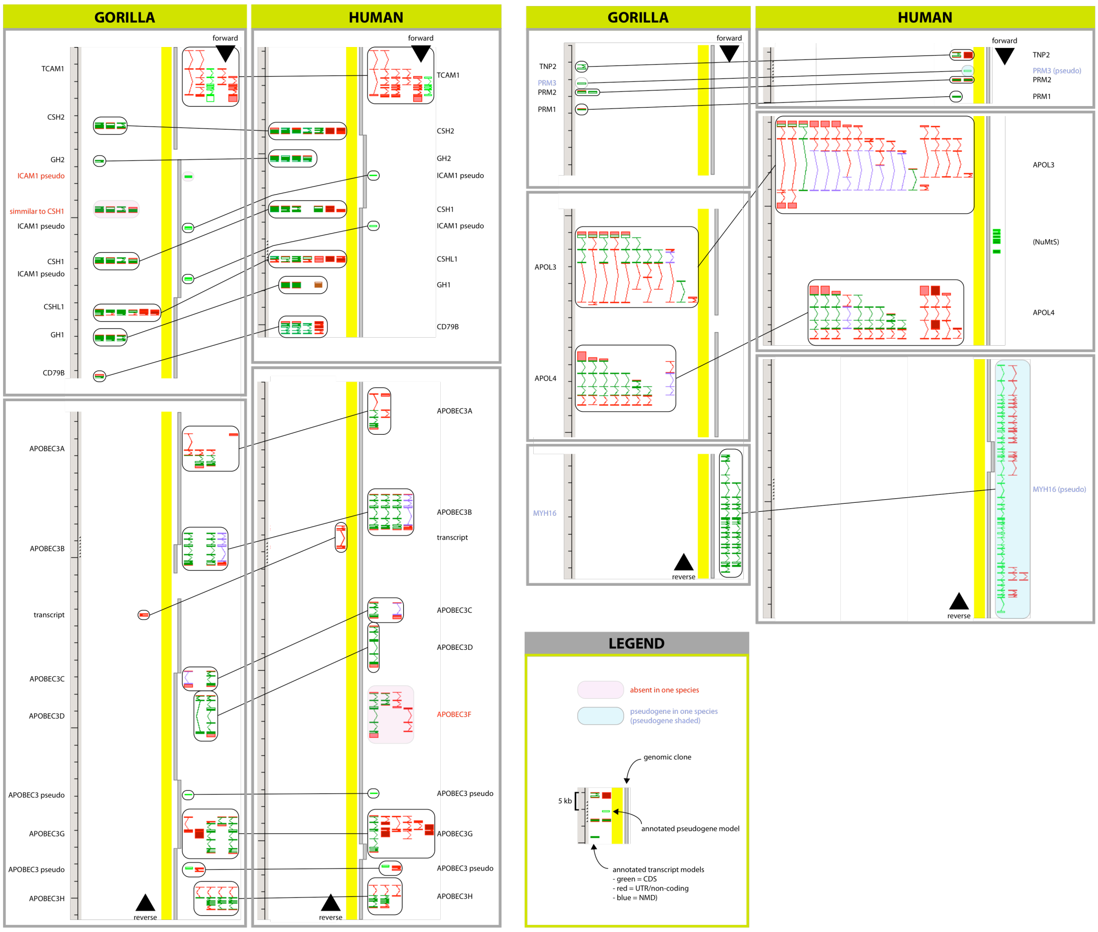


Fig SF9.1: Annotated fosmid sequences: The first two columns show the alignment for the Gorilla fosmid which maps to human chromosome 22. The second two columns show the alignment for the fosmid which maps to chromosome 16. Horizontal breaks between vertical boxes indicate missing data. The forward and reverse strand are indicated with black arrows. Lines between vertical boxes show homology between predicted gene models. Predicted genes or pseudogenes found in one species but not the other are coloured as indicated in the figure key.

Evolutionary analysis

The longest predicted transcript of each locus was used for evolutionary analyses. Each gene was BLASTed against the nr database on genebank to obtain orthologous genes from other primate species where available. In addition the orthologous genes from species with annotated genomes on Ensembl were downloaded. Sequences were aligned in MEGA 4.0 (Tamura, Dudley et al. 2007) and phylogenetic trees were build using a Neighbour Joining method and bootstrap of 500 replicates. Gene families such as the growth hormones (GH and CSH genes) and PRMs were analyses together.

The rate of molecular evolution was analysed using PAML (Yang 2007). We performed two tests. First, we tested whether there is evidence for positive selection acting at a proportion of sites across the entire tree using the site models (M2a vs. M2). Second, we tested whether there is evidence for a change in selection along the Gorilla branch compared to the background rate (either deceleration or acceleration) using the branch models. As alignments of gene families include many segments where sequence is present for one gene but not another, all PAML analyses were performed without removing ambiguous data so percentage of adaptively evolving sites under the site models may be off.

Growth Hormone cluster

Growth hormone genes have previously been shown to have duplicated multiple times within primates and their pattern of molecular evolution suggests a role for adaptive evolution of their protein coding genes (Wallis 2001; Revol De Mendoza, Esquivel Escobedo et al. 2004). However gene conversion events have been documented in previous studies which suggest unravelling their evolutionary history will be complicated (Revol De Mendoza, Esquivel Escobedo et al. 2004).

Phylogenetic trees of the growth hormones found with the fosmids, combined with data from the nr genebank database and ensemble confirm the presence of an additional Gorilla specific duplication event (Fig. SF9.1). However bootstrap values are low and there is some uncertainty over the exact relationship between different copies, particularly in the CSH cluster. Further more gene trees produced using the more divergent terminal end of the genes produce different topologies than trees based on the rest of the alignment which has stronger homology across genes.

The site model tests in PAML on the shared region show a proportion of sites (1.6%) in these genes have evolved adaptively ($\omega = 5.168$, LR = 8.813, $p = 0.012$). There is also some evidence of adaptive evolution at the terminal end of GH2-like genes (proportion of sites = 37.5%, $\omega = 2.854$, LR = 23.577 $p < 0.001$). However this result depends on using the gene tree topology obtained just using the terminal end, and when the gene tree for the whole coding sequence is used in the analysis significance is lost (LR = 1.716 $p = 0.424$), suggesting a complicated pattern of evolution of these genes.

Branch models comparing the dN/dS on the Gorilla branches to the rest of the tree suggest a significant slow down of evolutionary rate in Gorillas (Gorilla $\omega = 0.1667$, Background $\omega = 0.6514$, LR = 5.337 $p = 0.0209$), suggesting stronger purifying selection on the growth hormones during Gorilla evolution. However when just the sequences with GH2-like terminal ends are considered there is no evidence of a Gorilla slow down (Gorilla $\omega = 1.0505$, Background $\omega = 0.8640$, LR = 0.223, $p = 0.637$) again suggesting a complex pattern of molecular evolution at these loci.

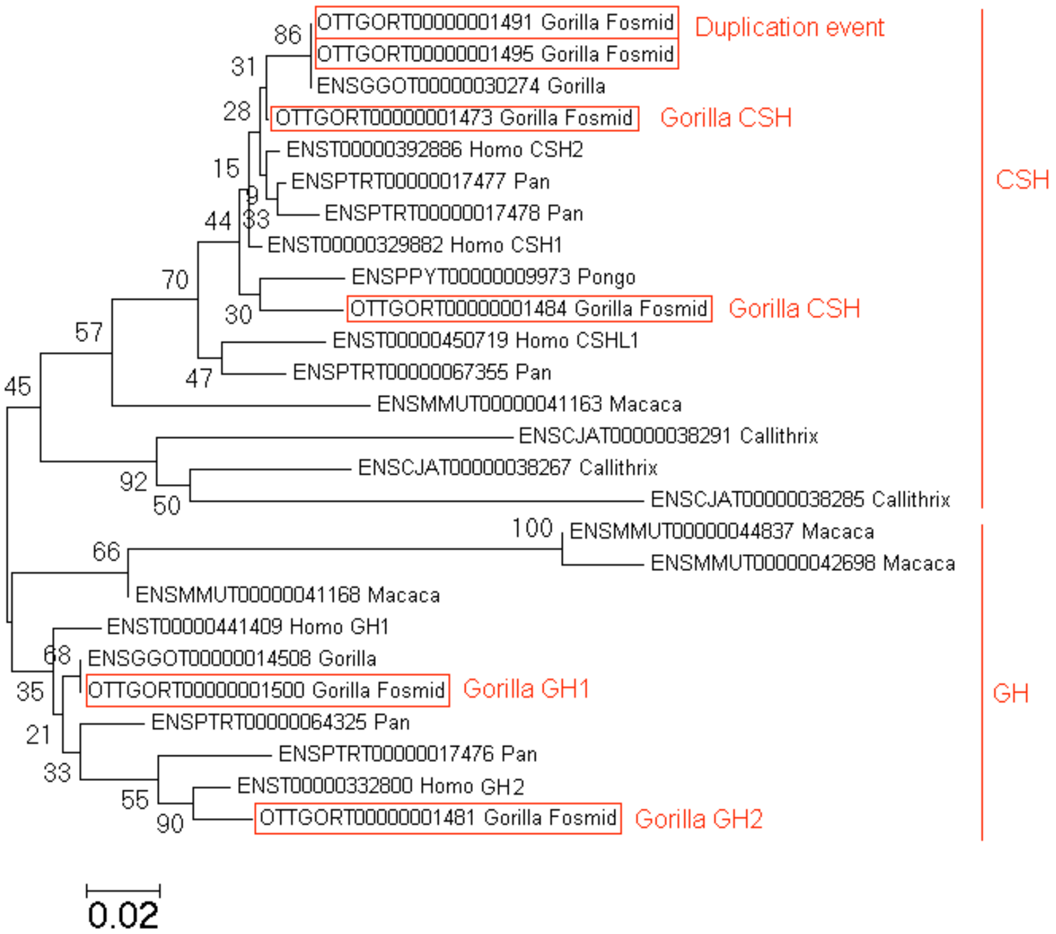

86
31
28
15
9
33
44
70
30
57
45
47
92
50
100
66
68
35
21
33
55
90
OTTGORT00000001491 Gorilla Fosmid
OTTGORT00000001495 Gorilla Fosmid
Duplication event
ENSGGOT00000030274 Gorilla
OTTGORT00000001473 Gorilla Fosmid
Gorilla CSH
ENST00000392886 Homo CSH2
ENSPTRT00000017477 Pan
ENSPTRT00000017478 Pan
ENST00000329882 Homo CSH1
ENSPPYT00000009973 Pongo
OTTGORT00000001484 Gorilla Fosmid
Gorilla CSH
ENST00000450719 Homo CSHL1
ENSPTRT00000067355 Pan
ENSMMUT00000041163 Macaca
ENSCJAT00000038291 Callithrix
ENSCJAT00000038267 Callithrix
ENSCJAT00000038285 Callithrix
CSH
ENSMMUT00000044837 Macaca
ENSMMUT00000042698 Macaca
ENSMMUT00000041168 Macaca
ENST00000441409 Homo GH1
ENSGGOT00000014508 Gorilla
OTTGORT00000001500 Gorilla Fosmid
Gorilla GH1
ENSPTRT00000064325 Pan
ENSPTRT00000017476 Pan
ENST00000332800 Homo GH2
OTTGORT00000001481 Gorilla Fosmid
Gorilla GH2
GH
0.02

Figure SF9.1: Gene tree of Growth Hormones (GH and CSH genes) showing Gorilla Fosmid sequences and Ensembl sequences (addition of further data doesn't greatly affect the tree) and the additional CSH duplication where the duplicate genes show high similarity to each other and an unnamed Gorilla CSH gene (presumably the same gene) annotated on Ensembl.

APOBEC cluster

Members of the APOBEC gene family generally cataylse the deamination cytosine to uracil. Rodents have one copy of APOBEC3 whereas humans have at least 6 copies, some of which have evolved adaptively apparently in relation to genome defense via RNA/DNA editing ((Sawyer, Emerman et al. 2004)). The phylogenies of the additional APOBEC sequences and the Gorilla Fosmid APOBECs show no evidence of Gorilla specific duplications (Fig. SF9.2). Several APOBEC3 genes are missing from the fosmids however they are predicted to fall in gaps in the fosmid sequence.

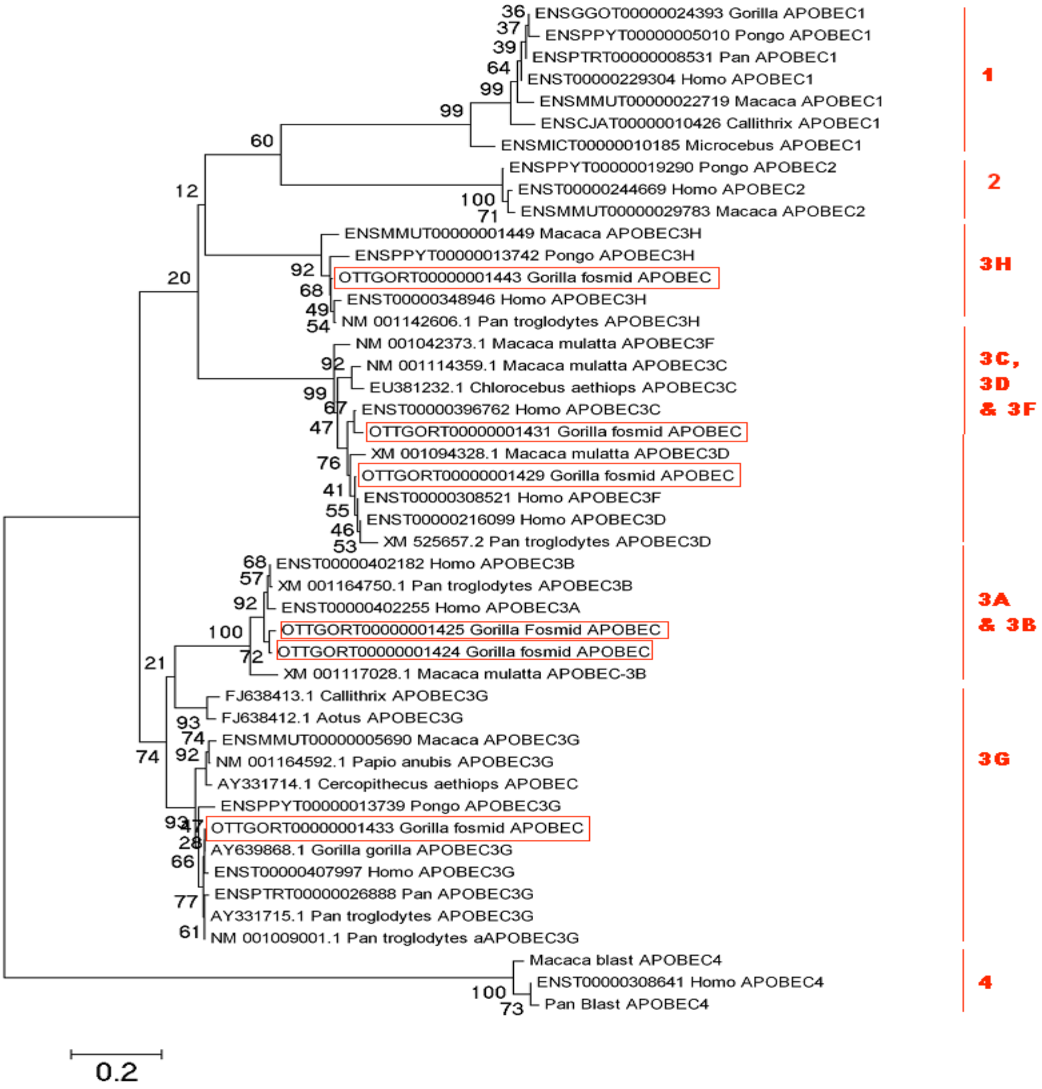

36
ENSGGOT00000024393 Gorilla APOBEC1
37
ENSPPYT00000005010 Pongo APOBEC1
39
ENSPTRT00000008531 Pan APOBEC1
64
ENST00000229304 Homo APOBEC1
99
ENSMMUT00000022719 Macaca APOBEC1
99
ENSCJAT00000010426 Callithrix APOBEC1
60
ENSMICT00000010185 Microcebus APOBEC1
ENSPPYT00000019290 Pongo APOBEC2
12
ENST00000244669 Homo APOBEC2
100
71
ENSMMUT00000029783 Macaca APOBEC2
ENSMMUT00000001449 Macaca APOBEC3H
ENSPPYT00000013742 Pongo APOBEC3H
92
OTTGORT00000001443 Gorilla fosmid APOBEC
20
68
ENST00000348946 Homo APOBEC3H
49
54
NM 001142606.1 Pan troglodytes APOBEC3H
NM 001042373.1 Macaca mulatta APOBEC3F
92
NM 001114359.1 Macaca mulatta APOBEC3C
99
EU381232.1 Chlorocebus aethiops APOBEC3C
67
ENST00000396762 Homo APOBEC3C
47
OTTGORT00000001431 Gorilla fosmid APOBEC
XM 001094328.1 Macaca mulatta APOBEC3D
76
OTTGORT00000001429 Gorilla fosmid APOBEC
41
ENST00000308521 Homo APOBEC3F
55
ENST00000216099 Homo APOBEC3D
46
53
XM 525657.2 Pan troglodytes APOBEC3D
68
ENST00000402182 Homo APOBEC3B
57
XM 001164750.1 Pan troglodytes APOBEC3B
92
ENST00000402255 Homo APOBEC3A
100
OTTGORT00000001425 Gorilla Fosmid APOBEC
72
OTTGORT00000001424 Gorilla fosmid APOBEC
21
XM 001117028.1 Macaca mulatta APOBEC-3B
FJ638413.1 Callithrix APOBEC3G
93
FJ638412.1 Aotus APOBEC3G
74
ENSMMUT00000005690 Macaca APOBEC3G
74
92
NM 001164592.1 Papio anubis APOBEC3G
AY331714.1 Cercopithecus aethiops APOBEC
ENSPPYT00000013739 Pongo APOBEC3G
93
47
OTTGORT00000001433 Gorilla fosmid APOBEC
28
AY639868.1 Gorilla gorilla APOBEC3G
66
ENST00000407997 Homo APOBEC3G
77
ENSPTRT00000026888 Pan APOBEC3G
AY331715.1 Pan troglodytes APOBEC3G
61
NM 001009001.1 Pan troglodytes aAPOBEC3G
Macaca blast APOBEC4
ENST00000308641 Homo APOBEC4
100
73
Pan Blast APOBEC4
1
2
3H
3C,
3D
& 3F
3A
& 3B
3G
4
0.2

Figure SF9.2: Gene tree of APOBEC genes

Site models show a strong signal of adaptive evolution across the APOBEC genes, with 13.2% of sites evolving with an average dN/dS of 2.761 (LR = 64.204, p <0.0001). However the rate of evolution along the Gorilla lineages does not differ from the background rate (Gorilla ω = 0.6765, background ω = 0.753 LR = 0.104 p = 0.747). The APOBEC4 cluster was not included in the PAML analysis as they appear to align in a different frame to the others. APOBEC3 genes were then realigned without APOBEC1,2 & 4, a new gene tree was built and the evolutionary rate of this cluster reanalysed. A site model test for positive selection just on the APOBEC3 cluster shows very strong evidence for adaptive evolution (LR = 128.189, p <0.0001) with 25.45% of sites evolving with an average dN/dS of 2.871.

PRM genes

Protamines replace histones in sperm cells during spermatogenesis and PRM1 has previously been found to have experienced positive selection in apes ((Wyckoff, Wang et al. 2000)). One copy of PRM 1-3 is found in the fosmids, with no evidence for duplications (Fig. SF9.3).

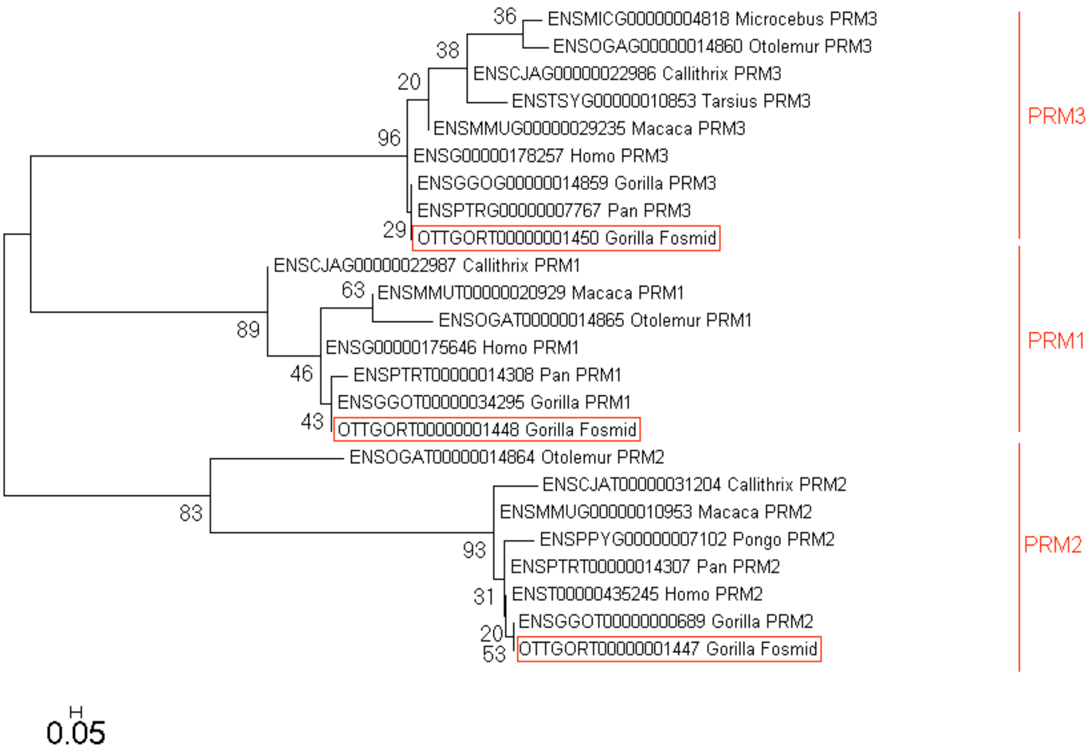

36
ENSMICG00000004818 Microcebus PRM3
38
ENSOGAG00000014860 Otolemur PRM3
20
ENSCJAG00000022986 Callithrix PRM3
ENSTSYG00000010853 Tarsius PRM3
ENSMMUG00000029235 Macaca PRM3
96
ENSG00000178257 Homo PRM3
ENSGGOG00000014859 Gorilla PRM3
ENSPTRG00000007767 Pan PRM3
29
OTTGORT00000001450 Gorilla Fosmid
PRM3
ENSCJAG00000022987 Callithrix PRM1
63
ENSMMUT00000020929 Macaca PRM1
ENSOGAT00000014865 Otolemur PRM1
89
ENSG00000175646 Homo PRM1
46
ENSPTRT00000014308 Pan PRM1
ENSGGOT00000034295 Gorilla PRM1
43
OTTGORT00000001448 Gorilla Fosmid
PRM1
ENSOGAT00000014864 Otolemur PRM2
ENSCJAT00000031204 Callithrix PRM2
ENSMMUG00000010953 Macaca PRM2
83
ENSPPYG00000007102 Pongo PRM2
93
ENSPTRT00000014307 Pan PRM2
31
ENST00000435245 Homo PRM2
ENSGGOT00000000689 Gorilla PRM2
20
53
OTTGORT00000001447 Gorilla Fosmid
PRM2
0.05

Figure SF9.3: Gene tree of PRM genes

Consistent with previous analyses (Wyckoff et al., 2002) we find significant evidence for positive selection acting on PRM1 (LR = 20.712, p <0.001) which has 44% of sites evolving with a dN/dS of 6.902. PRM2 and PRM3 do not appear to show any evidence for positive selection (LR = 0.355, p = 0837; LR = 0 p = 1.0 respectively). We find no evidence for a shift in selection pressure on any gene during Gorilla evolution, but this may be due to a small number of substitutions being involved. Wyckoff et al. also noted PRM1 is the only gene of the three to be 'present in sperm in its entirety'.

10. Transcriptome analysis

Analysis of transcriptome variation between species by RNA sequencing offers several advantages over previously used gene expression arrays that have suffered from the problem of designing probes in an unbiased way. Thus, RNA sequencing is likely to yield a more accurate quantification of gene expression levels. Additionally, sequencing provides information of transcription along the entire gene, enabling the analysis of differences in splicing between species. Thus, we sought to analyze the transcriptome of four primate species: gorilla, chimpanzee, bonobo, and human.

Total RNA was extracted from lymphoblastoid cell lines of one gorilla individual ("Murphy"), 2 chimpanzee, and 2 bonobo individuals. We sequenced each individual with paired end sequencing in one lane in an Illumina GAII sequencer. The read lengths and insert sizes are given in Table ST10.1. Additionally, we used RNA sequencing data of 8 human individuals (Montgomery, Sammeth et al. 2010).

Sample	Species	Lanes	Read length (bp)	Insert Size (bp)	Total reads (million)	Mapped reads (million)
GG013	*Gorilla gorilla*	4	1 lane: 2 x 54 3 lanes: 2x 76	200	65.4	29.6
EB176	*Pan trologdytes*	1	2 x 54	300	12.8	7.2
PTR8	*Pan trologdytes*	1	2 x 54	300	17.1	9.7
PC1582	*Pan paniscus*	1	2 x 54	200	40.2	22.7
PC1583	*Pan paniscus*	1	2 x 54	300	31.5	17.7
NA11994	*Homo sapiens*	1	2 x 36	200	19.3	10.0
NA12003	*Homo sapiens*	1	2 x 36	200	15.2	8.2
NA12045	*Homo sapiens*	1	2 x 36	200	26.9	14.1
NA12154	*Homo sapiens*	1	2 x 36	200	25.8	13.6
NA12812	*Homo sapiens*	1	2 x 36	200	25.1	13.1
NA12814	*Homo sapiens*	1	2 x 36	200	24.5	12.6
NA12874	*Homo sapiens*	1	2 x 36	200	21.5	10.0
NA12891	*Homo sapiens*	1	2 x 36	200	16.3	9.5

Table ST10.1: The RNA-sequencing dataset. The human data is from Montgomery et al.(Montgomery, Sammeth et al. 2010).

The sequence reads were mapped to the respective genomes (Ensembl 58; bonobo data mapped to the chimpanzee genome) with Maq 0.7.1, and Samtools 0.1.11 was used to extract information of sequence coverage. We used two different gene annotations to quantify reads mapping to exons, transcripts and genes. As a baseline, the Ensembl 58 annotations were used as standard gene models for each species, and we obtained the subset of these genes that were Ensembl 1:1 orthologs between all three species pairs – these are referred to as standard orthologous genes. However, in most species comparisons, we needed a more precise definition of orthology at exon level. To this end, we defined orthologous exons as sequence regions of 1:1 orthologous genes that are included the 5-way EPO alignment and annotated as exons in all the three species, with overlapping exons within a species pooled into one. The sum of orthologous exons per gene is referred to as exon-ortholog gene. For exon orthologs and exon-ortholog genes, those with more than 5% length difference between any of the species pairs have been excluded. Fig. SF10.1 shows statistics of mapping and quantification to different annotations. Table ST10.2 shows the numbers of genes and exons with read counts >=20 and >= 10, respectively, in different species. Altogether, we obtained data for 11849 nonredundant standard orthologous genes across species, 79013 orthologous exons, and 9746 exon-ortholog genes. These numbers are consistent with the expected number of genes expressed in the studied cell line. In subsequent analyses, only the genes and exons above these thresholds in at least one species were used, and in pairwise species comparisons, genes and exons under these thresholds in both species were excluded.

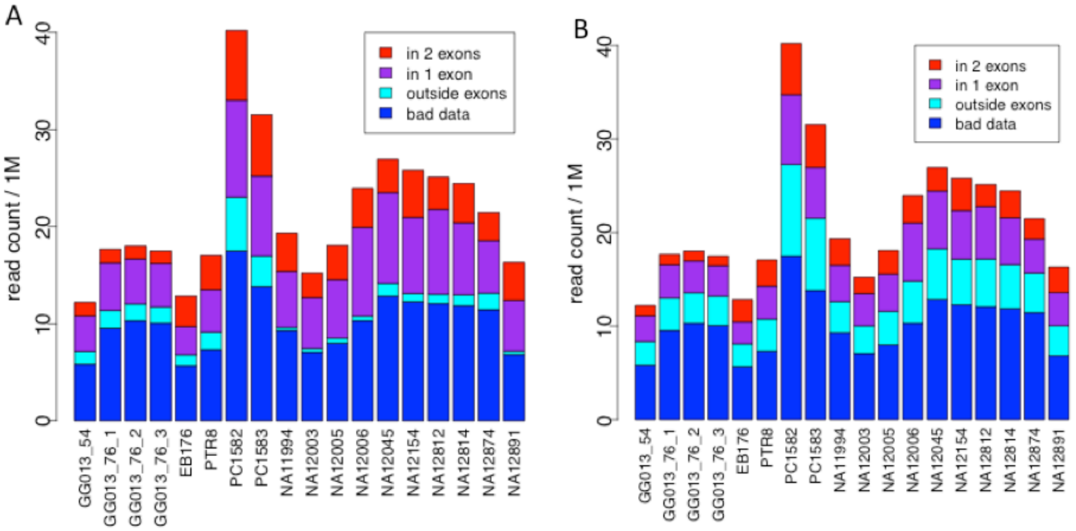


Figure SF10.1: Mapping statistics for RNA-sequencing lanes, using standard species-specific gene models (a) and orthologous exons (b).

	Human	Chimpanzee	Bonobo	Gorilla	Total
Standard gene models					
Total number of genes	51716	27166	27166	29217	NA
Genes with >= 20 counts	15778	11263	12640	11959	NA
Total number of orthologous genes	16511	16511	16511	16511	16511
Orthologous genes with >= 20 counts	10145	9484	10339	10000	11826
Total number of exons	562032	241850	241850	237229	NA
Exons with >= 10 counts	218338	85635	108872	62893	NA
Orthologous exons					
Total number of genes	14982	14982	14982	14982	14199
Genes with >= 20 counts	8274	8050	8934	8801	9746
Total number of exons	150187	150187	150187	150187	147686
Exons with > 10 counts	58232	51062	67874	59095	78857

Table ST10.2: Gene and exon counts for the different annotations. The genes > 20 and exons >= 10 counts are for means between lanes within species, and for any species for the total count. For orthologous exons, the exons and genes with >5% length difference between any species pair have been excluded from the total count.

We investigated the clustering of the samples by using the tools implemented in DESeq package. We normalized the counts in orthologous exons using variance stabilizing transformation, and calculated Euclidean distances between the lanes (Fig. SF10.2). Two human samples that appeared as outliers have been removed from the analysis, and the four lanes of data derived from a single gorilla individual have been pooled – they showed minuscule differences from each other despite one lane having a different read length. Notably, the high correlation between the bonobo samples with different insert sizes compared to sample pairs with the same size indicates that the variation in insert size is unlikely to be a major source of bias. Similar analysis based on exon-ortholog and standard ortholog gene quantification yielded similar results (data not shown).

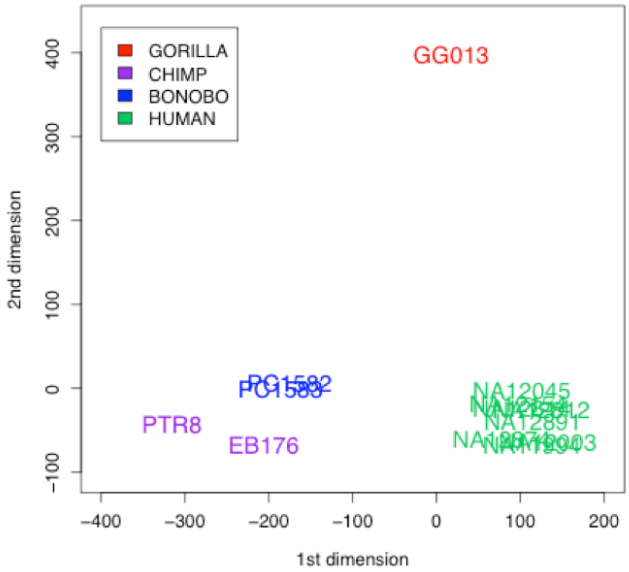


Figure SF10.2: A multidimensional scaling plot of Euclidean distances between samples based on normalized read counts in orthologous exons.

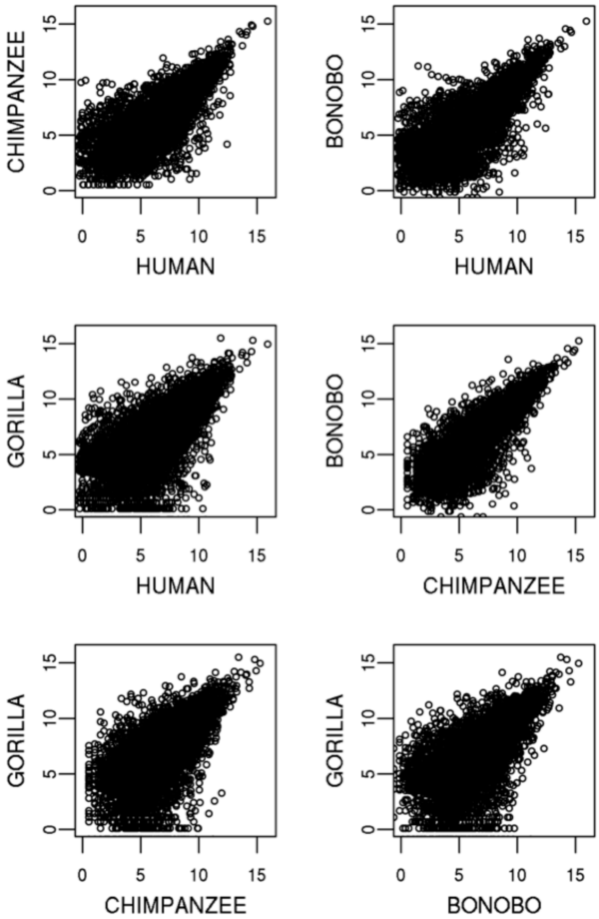


Figure SF10.3: Expression levels in exon-ortholog genes in the four species. The values correspond to median of log2 read counts in exon-ortholog genes per species.

Fig. SF10.3 shows mean expression levels of exon-ortholog genes between different species. As a measure of distance between expression values, we used the absolute distance between variance standardized counts in orthologous exons and exon-ortholog genes, first calculated between all sample pairs for each exon/gene and then obtaining distances between species pair as a mean of respective sample pairs. Fig. SF10.4 shows these distances of exon-ortholog genes for a multispecies comparison between gorilla, human and chimpanzee, enabling identification of genes with putatively gorilla-specific expression patterns.

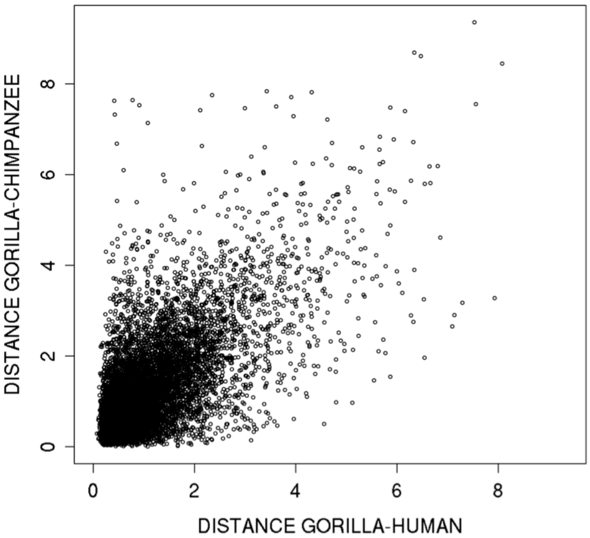


Figure SF10.4: Expression distances between exon-ortholog genes for gorilla-human and gorilla-chimpanzee.

Splicing variation was analyzed by comparing variation in expression distances between exons of a gene for the species pairs. In the absence of splicing differences between species, exons of a gene should show uniform distances between species, whereas splicing differences will increase the variation in exon distances. For all species pairs, these patterns were sought by first calculating expression distances of exons for all species pairs as the mean of expression differences of all corresponding pairs of individuals. The variance of square root normalized distances of all exons per gene was then used as a measure of splicing differences. We compared these differences between species to the diversity observed within humans by calculating the same statistics from expression distances between all the pairs of the eight human individuals. Fig. SF10.5 shows variance of exon distances plotted against gene distance for orthologous genes in the human-gorilla, human-chimpanzee and gorilla-chimpanzee comparisons, as well as the within-species human-human comparison. In each case, horizontal and vertical lines show the maxima and 99th percentiles of the human distribution. Fig SF10.6 illustrates gorilla-specific splicing in the *SQLE* gene, involved in steroid metabolism. From the numbers in Table ST10.3, averaging over the three inter-species comparisons, we calculate that 7% of genes show significant splicing variation at the 1% level, based on the distribution within human. However we note the caveat that differences in data quality and sample size may contribute to the differences in these numbers.

	Human-gorilla	Human-chimpanzee	Chimpanzee-gorilla	Human-human
Genes with exon variance > 99th centile of human distribution	359	214	942	58
Total genes compared	5178	4829	5171	5884

Table ST10.3. Quantities in the comparison of splicing differences between the great apes.

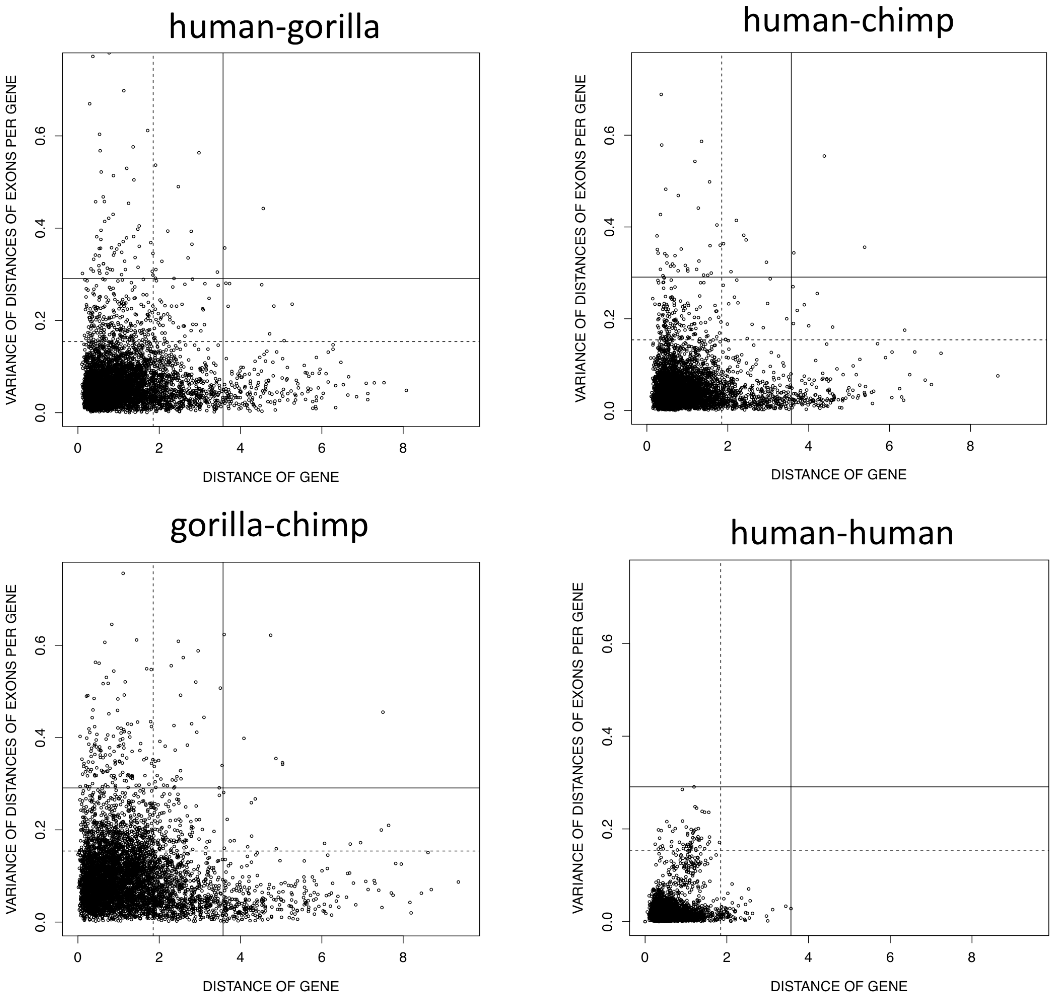


Figure SF10.5: Variance of exon distances within each gene plotted against gene distance for orthologous genes in the human-gorilla, human-chimpanzee, gorilla-chimpanzee and human-human comparisons. In each case, horizontal and vertical lines show the maxima and 99th percentiles of the human distribution.

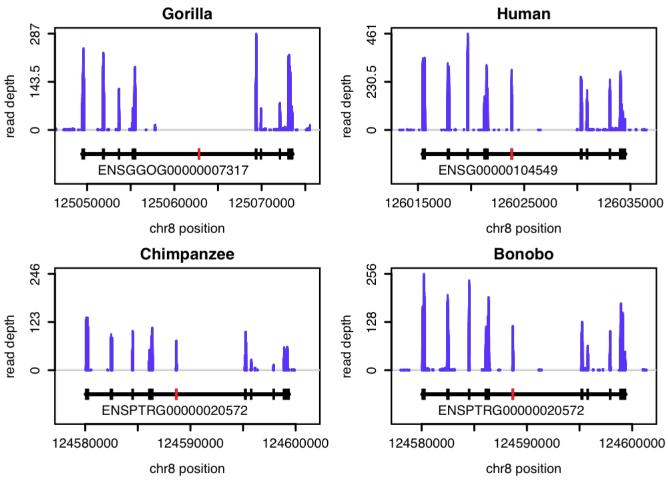


Figure SF10.6: RNA sequence coverage in the squalene epoxidase gene (SQLE) in the gorilla, human, chimpanzee and bonobo, with all data pooled within species. The exon highlighted in red is not expressed in the gorilla, with the preceding intron containing a 2.9 kb insertion in this species**.**

Next, we analyzed whether expression differences between species correlate with genetic differences. First, we studied whether incomplete lineage sorting in orthologous genes is correlated to expression patterns, which is expected in some of the cases when ILS is driven by selection. We compared the species tree obtained from the expression distances to the genetic trees for each gene (most frequent pattern per gene for alignment filtered data). In total, this analysis included 5577 genes with both expression and genetic data. Altogether, we observed a significant enrichment of having an identical tree pattern in genetic and expression data (p = 0.026 based on 100 000 permutations). This enrichment was not, however, evenly distributed (Table ST10.4): especially for genes with a genetic tree of humans as the outgroup, the expression data shows a clear underrepresentation of the normal species tree, and a strong enrichment for an ape-specific expression pattern, suggesting that especially in these cases a proportion of the ILS correlates with selection.

	genetic tree		
expression tree	G.CH	C.GH	H.CG
G.CH	1473 (+2.8%)	372 (-2.8%)	204 (-12.6%)
C.GH	1430 (-1.2%)	389 (+0.7%)	250 (+6.1%)
H.CG	997 (-2.3%)	281 (+3.1%)	181 (+9.0%)

Table ST10.4: Correspondance between phylogenetic trees of genes based on genetic and expression data. The percentages indicate difference from the null. The species abbreviations are human (H), chimpanzee (C), and gorilla (G).

We also analyzed whether genetic divergence correlates with expression divergence. We calculated the genetic identity for the coding regions of standard gene models between the three species pairs, and compared these to expression distances calculated for exon-ortholog genes. For most species pairs we observed a small but significant correlation (Table ST10.5), suggesting that genes with highly similar coding regions may be more likely to have similar expression profiles. This analysis may however be sensitive to different qualities of the genome sequences and the gene annotation in the apes relying on human annotation. A more consistent pattern of correlation is indeed observed when expression distances are compared to median Gerp conservation scores of genes based on genomes of 33 amniote species (Table ST10.5).

	Correlation between expression distance and pairwise sequence identity		Correlation between expression distance and amniote sequence conservation	
	rho	p	rho	p
human-chimpanzee	-0.039	3.4E-04	-0.034	1.8E-03
chimpanzee-human	-0.026	0.016	-0.055	8.0E-07
human-gorilla	0.004	0.735	-0.043	4.9E-05
gorilla-human	-0.037	4.1E-04	-0.079	9.6E-14
chimpanzee-gorilla	0.020	0.059	-0.056	1.6E-07
gorilla-chimpanzee	-0.066	3.7E-10	-0.065	1.2E-09

Table ST10.5: Correlation between genetic and expression divergence. The name listed first for each species pair denotes the species whose annotation is used for calculating the sequence identity to the other species, or the median Gerp conservation scores, and the table shows the Spearman correlation between these and the expression distance for exon-ortholog genes.

11. ChIP-seq analysis

The EB(JC) western gorilla LCL line was used for gorilla ChIP-seq assays as recently described (Schmidt, Wilson et al. 2009). Briefly 1x10^8 cells were cross-linked with 1% formaldehyde and CTCF-bound DNA immunoprecipitated with a CTCF antibody (Millipore, 07-729). Immunoprecipitated DNA was end-repaired, A-tailed and single-end Illumina sequencing adapters ligated before 18 cycles of PCR amplification. 200-300 bp DNA fragments were selected and 36 bp reads sequenced on an Illumina Genome Analyser II according to manufacturer's instructions. The experiment was performed in duplicate.

We used publicly available ENCODE data for the CEU European cell lines GM12878, GM19238 and GM12891 (McDaniell, Lee et al. 2010). The raw reads were aligned to the reference genomes (human GRCh37 or gorilla gorGor2) using Bowtie (Langmead, Hansen et al. 2010) with the parameters "-n 2 -m 3 -k 1 --best", allowing two mismatches and excluding reads mapping to over 3 locations in the genome. Reads were filtered for y-chromosome and mitochondrial DNA. Peaks were called on all datasets using MACS with default parameters and matched input (Zhang, Liu et al. 2008). The intersection of called peaks from replicates was used for further analysis and thus represents a stringent subset of peaks.

CTCF binding events in both species were aligned using the Enredo-Pecan-Ortheus 6-way primate alignment available in Ensembl release. 60. 97% of human and 94% of gorilla CTCF binding events are contained within the primate-EPO alignment, and constitute the analysis space for all inter-species comparisons. To determine the CTCF sequence motifs, we first performed *de novo* motif discovery with NestedMica (Down and Hubbard 2005) using the parameters ”minLength 10 -maxLength 35 -numMotifs 6 -revComp” within a 50 bp window around the peak summit for the top 1000 human peaks ranked by MACS score. We then used the longest found motif to scan the entire peak regions for a motif match using the NestedMica utility nmscan at a score cutoff of -15.

We analysed the intersect of two replicates for the three human ENCODE cell lines to obtain a measure of technical and inter-individual variation. While ~30 000 CTCF bound regions are shared among all three individuals, roughly 10 000 are unique to each human cell line (Fig. SF11.1A). Sites that are common to all three individuals are much more likely to be shared with gorilla (~ 60% of the individual-invariant sites versus < 20% of either one or two-individual specific) (Fig. SF11.1B). Previous studies in human have estimated that up to 25% binding site differences occur within species (Kasowski, Grubert et al. 2010; McDaniell, Lee et al. 2010).

For the inter-species comparison, we used a single cell line (GM12878), for which the fraction of CTCF binding sites associated with CpG islands was in the range of our gorilla data, as well as several other publicly available CTCF datasets (data not shown). Comparison of all human and gorilla CTCF bound regions revealed that species-specific binding events are generally similar to shared ones, with small differences in ChIP enrichment and motif content (Fig. SF11.1C-D). Species-specific regions tend to be smaller with respect to peak width, number of tags per region and have a slightly lower occurrence of the CTCF motif.

The association of CTCF sites within CpG regions of the genomes was determined using existing CpG island annotations available in Ensembl release 60 for both species. In order to determine the genetic mechanisms underlying differences in CTCF binding, regions were classified on the basis of species occupancy (human-specific, shared and gorilla-specific) and CpG island association (non-CpG, shared CpG, species-specific CpG). For these classes we identified potentially bound CTCF motif instances and projected these onto the respective genome. We then quantified the proportion of indels, disruptions (gaps > 4 bp), substitutions and unchanged motifs.

We analysed the distribution of binding events in the genome by plotting inter-peak and inter-gene distances in three classes: gorilla-specific, shared and human-specific binding events. We found that a higher fraction of species-specific peaks are located at relatively small (up to 2kb) distances from each other compared to shared peaks (data not shown) and that the fraction of species-specific binding sites with an inter-peak distance of less than 400 bp is significantly higher than for the shared binding events (p=5.251702*10-114 Fisher's one-sided exact test).

To test whether there is a high-level association between expression changes and CTCF binding, we mapped the number of species-specific, and shared binding sites per orthologous gene pair. We found a small, but significant correlation between the number of species-specific peaks and Euclidian expression distance per gene – Spearman rho 0.05 and p=9.156*10-07. We plotted the INSR locus that shows high expression differences as well as differential CTCF binding between human and gorilla by using the Ensembl Genome Browser in the multiple alignment primate EPO view, showing 1-1 positioning of one lane of CTCF ChIP-seq, as well as one lane of RNA-seq data in both human and gorilla.

We found a small correlation (rho = 0.05, p = 9.1 ´ 10-7) between gene expression difference and the number of species-specific CTCF binding events per gene. Differential CTCF binding accompanying changes in gene structure and transcription (summarized in Table ST11.1) is illustrated at the insulin receptor (INSR) locus (Fig SF11.2), where regulatory changes might contribute to species-specific biology of glucose uptake.

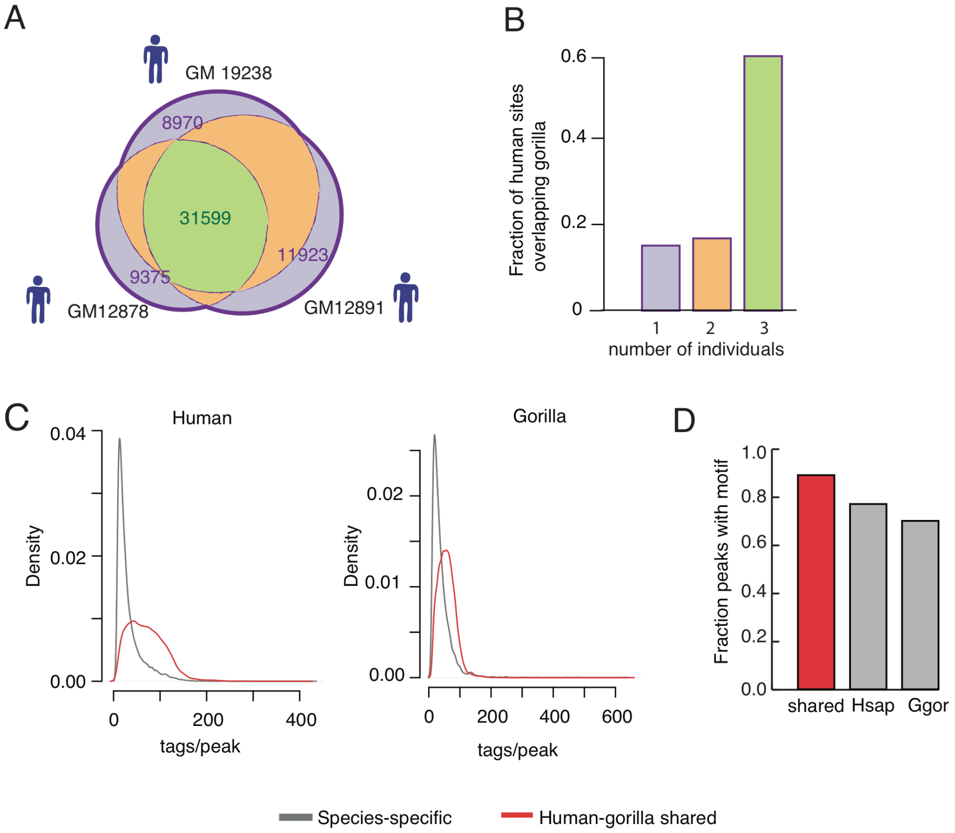


Figure SF11.1: CTCF ChIP-seq analysis in human and gorilla. (A) Three-way human CTCF overlaps. Binding events in three unrelated human ENCODE cell lines (GM12878, GM19238, GM12891) were overlapped to reveal numbers of shared and individual-specific bound regions. (B) Human CTCF binding events bound in gorilla. Regions that overlap gorilla specific for one, two or three human individuals are shown. (C) CTCF ChIP enrichment in species-specific and shared categories. (D) Fraction of human-specific, human-gorilla shared and gorilla-specific peaks that contain at least one CTCF motif.

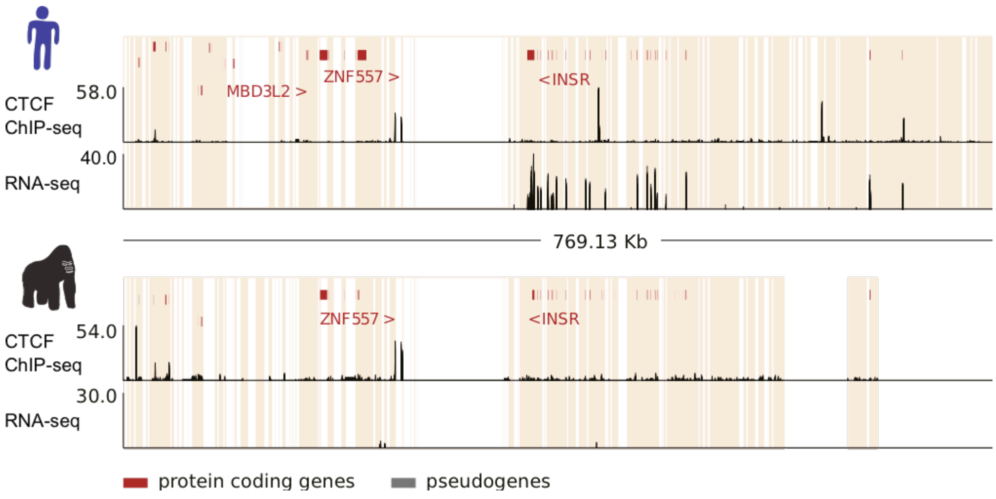


Figure SF11.2: CTCF ChIP-seq and RNA-seq expression data in human and gorilla at the *INSR* locus. Brown blocks represent syntenous regions in the primate EPO alignment; Ensembl gene annotation is shown in red

12. Genomic comparison of the Eastern and Western gorilla species

To explore genetic variation within the *Gorilla* genus, we mapped reduced representation sequence data for the female Western Lowland gorilla EB(JC) and the male Eastern Lowland gorilla Mukisi (see ”Genome Sequencing” above) to the gorilla reference assembly. Excluding sites covered by less than five reads, this produced an average coverage depth of 111x in EB(JC) and 71x in Mukisi within short regions distributed genome wide, representing ~1.2% of genomic material in each individual, and with highly concordant representations: e.g. of sites covered by more than five reads in EB(JC), 92% were similarly covered in Mukisi. Moreover, coming from the same sequencing run, reads from each sample shared identical base-calling error characteristics.

This data, along with the whole-genome Illumina sequence data for Kamilah, was mapped to the gorilla reference assembly using BWA (Li and Durbin 2010). Variants were called using SAMTOOLS (Li, Handsaker et al. 2009), filtering sites to include only those with depth > 20 and < 90, and excluding those with consensus quality less than 20. The rates of heterozygous and homozygous variant calls were calculated and shown in Table 2 in the main paper.

We note that the hom:het ratio for EB(JC) (0.56) is consistent with her coming from the same population as the assembled individual (Kamilah), as the expected ratio in this case is 0.5. To see why, consider the unrooted tree corresponding to the alignment of a diploid sample to a reference sequence, a tree with three leaves and three branches. Under the assumption of infinite sites each variant is caused by a mutation on just one of the branches, of which two give rise to heterozygous variants in the sample and only one (the branch leading to the reference) gives a homozygous non-ref variant. Thus we expect twice as many hets as homs whenever the sample chromosomes and the reference are interchangeable - i.e. when they are from the same population.

Recall that in a panmictic population the rate of heterozygosity $\theta = 4N\mu g$ where N is the population size, μ is the mutation rate, and g is the generation time. Using the values in Table 3 for EB(JC) and Mukisi ($\theta_W = 0.178\%$ and $\theta_E = 0.076\%$ respectively), and assuming $\mu = 0.6 \times 10^{-9}$ per year per site and g = 20 years, we obtain $N_E = 15{,}800$ and $N_W = 37{,}000$. Accounting for the different mutation rate used, these numbers are similar to those reported in (Thalmann, Fischer et al. 2007).

The data for EB(JC) and Mukisi was then mapped to the human reference sequence (GRCv37), and variants again called using samtools with filtering as above. Sites in repeat-masked regions of the human reference or in annotated segmental duplications were also excluded. Sites which passed these restrictions in both individuals, and in which one variant (non-human) allele was called in one or both individuals, were classified by genotype and counted as shown in Table ST12.1.

EB(JC) genotype	Mukisi genotype	Sites in human alignment	Proportion
0/0	0/1	1919	0.02328
0/0	1/1	2656	0.03222
0/1	0/0	5284	0.06409
0/1	0/1	1401	0.01699
0/1	1/1	1794	0.02176
1/1	0/0	1727	0.02095
1/1	0/1	595	0.00722
1/1	1/1	67065	0.81349

Table ST12.1: Variant sites in alignment of EB(JC) and Mukisi data to human. Genotypes are encoded with 0 representing a match to the human allele and 1 a mismatch.

We can show that the distribution of site counts in Table ST12.1 is not compatible with a clean split between the two species. In particular, consider the sites with pattern 0101, 1101 and 0111 (where e.g. 0111 means genotype 0/1 in the Western individual and 1/1 in the Eastern): in each case, under a clean split model the variant mutation must have occurred within the ancestral population (discounting recurrent mutations). Looking back in time, let the probability of two chromosomes failing to coalesce at a locus before reaching the ancestral population be α in the Western population and β in the Eastern. Let θ be the scaled mutation rate in the ancestral population, and let $q = \theta L$, where L is the number of positions across the genome at which we are able to call genotypes in both individuals, given the filtering constraints. There are then three scenarios to consider with respect to these sites:

1. With probability $\alpha\beta$, all four chromosomal lineages remain separate back to the ancestral population. Then, if we assume a constant effective population size before the split, we can use the result known both from population genetics and coalescent theory that in a set of samples the expected rate of segregating sites exhibiting n derived alleles is θ/n (e.g. (Ewens 2004; Hein, Schierup et al. 2005)). Thus in this case the number of sites with three derived alleles will be $q/3$, of which half will be 0111 and half 1101. The number of sites with two derived alleles will be $q/2$, of which 2/3 will be 0101.

2. With probability $(1 - \alpha)\ \beta$, both Western chromosomes coalesce before the ancestral population but the Eastern chromosomes do not, meaning that three lineages exist at the time of speciation. Then the number of sites with two derived alleles present in these three lineages will be $q/2$, of which 2/3 will result in 1101 sites and 1/3 0011 in the four present-day chromosomes. Sites with one derived allele at speciation will give 1100 or 0001 configurations.

3. With probability $\alpha\ (1 - \beta)$, both Eastern chromosomes coalesce before the ancestral population but the Western chromosomes do not; again three lineages exist at the time of speciation. The number of sites with two derived alleles present in these three lineages will be $q/2$, of which 2/3 will result in 0111 sites and 1/3 1100. Sites with one derived allele at speciation will give 0011 or 0100 configurations.

Thus we can write the following expressions for the expected numbers of 0101, 1101 and 0111 sites:

$$E(n_{0101}) = \frac{1}{3} q\alpha\beta$$

$$E(n_{1101}) = \frac{1}{6} q\beta(2-\alpha)$$

$$E(n_{0111}) = \frac{1}{6} q\alpha(2-\beta)$$

from which we get estimators for α, β:

$$\hat{\alpha} = \frac{2n_{0101}}{2n_{1101} + n_{0101}}, \quad \hat{\beta} = \frac{2n_{0101}}{2n_{0111} + n_{0101}}$$

Now since α, β are probabilities, they must both lie in the interval [0, 1], meaning we must have $n_{0101} \leq 2n_{0111}$ and $n_{0101} \leq 2n_{1101}$ for consistency with a clean split model. In our data we find n_{0101} / n_{1101} = 2.35, with a 95% CI of (2.12, 2.58). In other words, there are too many shared heterozygous sites in the data for compatibility with a model in which the Eastern and Western Lowland gorilla populations separated with no subsequent genetic exchange.

Effect of model assumptions

The calculation makes two assumptions about the data, namely that the infinite sites approximation is valid (we ignore recurrent mutations) and that the ancestral population was unchanging in effective size. The former is reasonable in this context as we have only counted sites which are bi-allelic across all four chromosomes.

To address the impact of possible variation in ancestral population size on our conclusions, we have carried out several coalescent simulations. Figure SF12.1 shows the results of these.

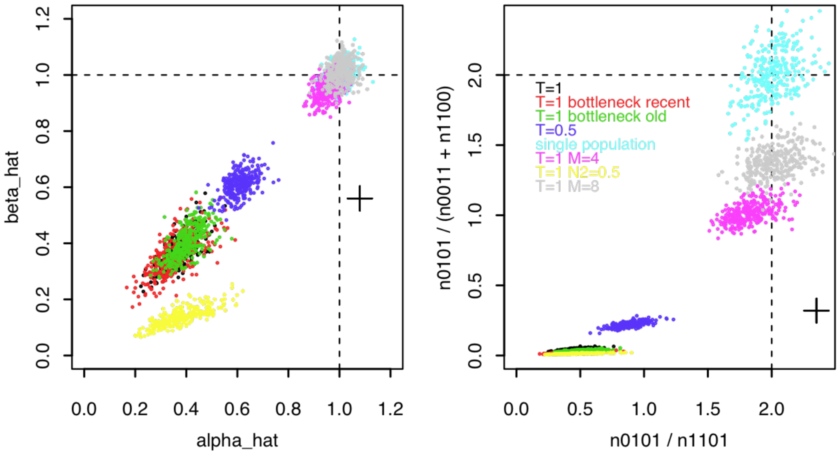


Figure SF12.1. Coalescent simulations of a population split under different demographic models. Each dot represents a simulation of 5000 loci under the demographic model represented by its colour, indicated in the legend on the right-hand panel. Here T is the split time between the species, given in units of 2Ne (N_e the effective size of the ancestral and post-separation populations, except where otherwise indicated); 'bottleneck recent' occurs between t = 1.1 and 1.6, during which the ancestral population is reduced to $0.2N_e$; 'bottleneck old' is the same but occurs between t = 2.0 and 2.5. N2 is the effective size of the 'eastern' population. We also simulated the effects of migration between the species after separation, represented by M in units of $4N_em$, where m is fraction of the total population migrating each generation. **left,** alpha_hat (x-axis) and beta_hat (y-axis) are estimators of the probabilities of two lineages avoiding coalescence within the west and east gorilla populations respectively. Under a clean split model these are expected to both lie between 0 and 1. **right**, the x-axis n_{0101}/n_{1101} should be less than 2 given a clean split, while the y-axis (n_{0101} / (n_{1100} + n_{0011})) is the ratio of shared heterozygous sites to 'fixed' or homozygous differences. This has an expected value of 2 for a single population with no split, but will be less than 2 when a split has occurred, introducing extra 1100 and 0011 sites. The cross marks where the gorilla data falls on these plots.

Clearly, none of the clean split scenarios give results close to the observed data points. In particular, the two bottleneck scenarios have negligible effect on the estimators, seen clearly on the left-hand panel.

We do see simulated results closer to the data when migration has occurred (which invalidates the clean split assumption), and the right hand panel shows that the y-axis distinguishes between recent and older splits even when migration is nonzero. However we haven't attempted to fit by eye set of parameters which matches the gorilla data here. The space is very large, and there are several other parameters one could choose to vary, such as N1, N2 and asymmetry or time variation in migration.

These simulations show that our criteria for the absence of gene flow are tolerant of time variation in the ancestral population size. Note also that selection, which corresponds to variation in N_e between loci, can be represented simply by averaging or combining models with different N_e parameters, since the loci in our dataset are independent. Therefore this also should also not effect our inference of gene flow between species.

Sequence divergence estimate

We conclude that speciation between eastern and western gorillas occurred as an initial divergence followed by a period of limited genetic exchange. The timing of this initial divergence is difficult to estimate because it is not well defined. Indeed, some degree of separation or substructure with exchange could have existed for an indefinite time. Moreover this exchange could have occurred continuously, at constant or fluctuating levels, or it could have been intermittent, perhaps even comprising just one short significant episode. It may also have been asymmetric between the two species. The inference of historical gene flow between populations using present-day genetic data constitutes a challenging inverse problem.

We can estimate the average sequence divergence d_{EW} between the species. First we note that

$$d_{EW}L = \frac{1}{2}\left(n_{0100} + n_{0111}\right) + n_{1100} + \frac{1}{2}\left(n_{0001} + n_{1101}\right) + n_{0011}$$

and L can be eliminated by scaling to the human-gorilla sequence divergence, via the total height of the tree, as follows.

For any phylogeny in which the branches correspond to segregating sites as shown in Fig. SF12.3, we can estimate the total mean height of the tree (in units of number of mutations) by averaging over the paths from each leaf to the root:

$$H = \frac{1}{k}\sum_{i=1}^{k-1} iS_i$$

where k is the number of leaves (chromosomes) and S_i is the number of sites with i derived alleles.

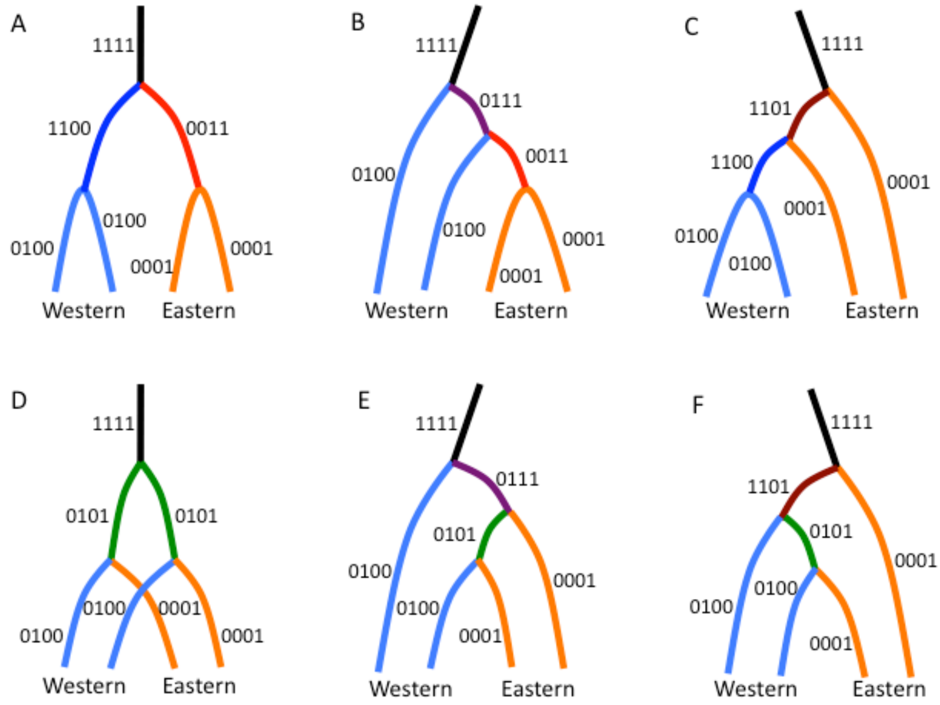


Figure SF12.3: **Ancestral tree topologies relating four chromosomes**: two from the Western Lowland gorilla EB(JC) and two from the Eastern Lowland gorilla Mukisi. In total there are 15 possible rooted tree topologies describing the ancestry of the four chromosomes at each of the sites counted in Table S3 (with human as an outgroup); however since our variant calls are unphased, we do not distinguish between the chromosomes within each individual, leaving us with just six topologies to consider. In each tree, branches are labelled and coloured according to the present-day genotypes which would result from a variant mutation on that branch (under an infinite sites assumption).

In this case we have

$$H = \frac{1}{4}\left(n_{0100} + n_{0001} + 2\left(n_{1100} + n_{0011} + n_{0101}\right) + 3\left(n_{0111} + n_{1101}\right)\right)$$

where e.g. n_{1100} is the number of sites with genotype 1/1 in EB(JC) and 0/0 in Mukisi (and recall that 1/0 sites in either individual are counted as 0/1 since we only have genotypes). We can calibrate this to the sequence divergence d_{HG} between human and gorilla by noting that

$$d_{HG}L = H + n_{1111}$$

so that

$$\frac{H}{d_{HG}L} = \frac{H}{n_{1111} + H}$$

Using the site calls for the alignment of EW_JC and Mukisi data to human we calculate $H/d_{HG}L$ = 0.088 with a 95% bootstrap confidence interval (CI) of 0.0865-0.0895, and $d_{EW}L/H$ = 1.42 (CI 1.40-1.44).

Then we can write

$$d_{EW} = \frac{d_{EW}L}{H}\frac{H}{d_{HG}L}d_{HG}$$

Previously we found d_{HG} = 0.017 (Table ST3.3), which gives d_{EW} = 0.0021. Taking μ = 0.6 x 10^{-9} bp^{-1}y^{-1} gives a sequence divergence time $d_{EW}/2\mu$ = 1.75 Mya.

Isolation-migration model

In order to estimate the speciation time and amount of gene flow between the Eastern and Western Lowland gorilla subspecies we implemented a two species isolation-with-migration (IM) model (Nielsen and Wakeley 2001) (Fig. 4c in the main paper). For two diploid individuals, at any given site there are 100 possible genotype combinations $x_1x_2{:}y_1y_2$, x_1, x_2, y_1, $y_2 \in$ (A, C, G, T). Our approach compares the counts of each of these combinations in the data for EB(JC) and Mukisi with probabilities calculated by summing over all possible topologies for the four lineages. In each topology mutations are accounted for by integrating out the coalescent prior with a nucleotide substitution model; we assumed a general time reversible substitution model with parameters estimated from the data.

The full IM model has 6 parameters: The population sizes for the two gorilla species and their ancestor are N_W, N_E and N_A and the split time is τ years ago. Migration events from one population to the other and backwards in time have rates M_{WE} and M_{EW} per generation. Migration rate parameters have units of migrations per mutation event such that $m = M / (\mu g)$. Rates of the two kinds of migration events are denoted by m_{WE} and m_{EW}. The scaled parameters are $\theta_W = 4N_W\mu g$, $\theta_E = 4N_E\mu g$, $\theta_A = 4N_A\mu g$, $m_{WE} = M_{WE} / (\mu g)$, $m_{EW} = M_{EW} / (\mu g)$ and $T = \tau\mu$.

However, we observed identifiability issues in the full model using maximum likelihood estimation, in the sense that a very wide range of parameters were compatible with the data, including almost arbitrarily ancient speciation times. Therefore we explored a constrained model in which $\theta_W = \theta_A = 0.178\%$ and $\theta_E = 0.076\%$ (effectively fixing population sizes at their present-day values, and with ancestral population size equal to that of the Western lowland population) and symmetric migration $m_{WE} = m_{EW} = m$. We therefore only have two unknown parameters in this model, namely the split time and the migration rate. In Fig. 4d we show the log-likelihood for various values of the speciation time and migration rate, with a maximum likelihood solution at $\tau = 500$ kya, and a migration rate of 0.2 migration events per generation, assuming a 20-year generation time and a mutation rate of 0.6×10^{-9} per bp per year. Nevertheless we reiterate that additional degrees of freedom, such as varying population sizes or asymmetry in gene exchange, allow a much broader range of solutions.

Protein coding differences between gorilla species

Additional capillary (Sanger) sequencing was undertaken to check selected protein coding variants differentiating the western gorillas from the eastern individual Mukisi, either as non-synonymous SNPs or stop codons.

Chr:Position (gorGor3)	*Gg* Ref	Alt	WESTERN						EASTERN	
			Kamilah (Reference)		Kwanza		Murphy (Gg013)		Mukisi	
			Illumina	Sanger	Illumina	Sanger	Illumina	Sanger	Illumina	Sanger
Chr19:14955128	C	T	CC	C	CC	C	No Call	C	TT	T
Chr19:14958507	T	C	TT	T	TT	T	No Call	T	CC	C
Chr19:14958517	A	G	AA	A_7G	AA	A_7G	No Call	A_7G	GG	A_6GG
Chr19:14958540	A	G	AA	A	AA	A	No Call	A	GG	G
Chr19:14958570	A	G	AA	A	AA	A	No Call	A	GG	G

Table ST12.3. SNPs variants in EMR3.

Chr:Position (gorGor3)	Gene	*Gg* Ref	Alt	WESTERN						EASTERN	
				Kamilah (Reference)		Kwanza		Murphy (Gg013)		Mukisi	
				Illumina	Sanger	Illumina	Sanger	Illumina	Sanger	Illumina	Sanger
Chr19:37077800	*CLC*	G	A	GG	G	GG	G	No Call	G	AA	A
Chr20:43075776	*SEMG2*	G	T	GG	G	GG	G	No Call	G	TT	T

Table ST12.4. Loss of function variants in eastern gorillas in the genes SEMG2 associated with sperm competition and the Charcot-Leyden crystal protein (CLC)/galectin-10 gene associated with inflammation.

13. Genome sequence duplications

We detected genomic duplications in the Western Lowland gorillas Kwanza (Ventura et al. submitted) and Kamilah using the whole-genome shotgun sequence detection (WSSD), an assembly independent method which identifies regions greater than 20 kbp in length with a significant excess of read depth within 5 kbp overlapping windows (Bailey et al. 2002). WSSD analysis was performed using whole genome Illumina reads from both gorillas, following correction methods specific to next-generation sequencing data (such as biases in the GC distribution) as previously described (Alkan et al. 2009). The reads were mapped within an edit distance of 4 to a repeat-masked version of the human genome using MRFAST, an algorithm that tracks all read map locations allowing read-depth to be accurately correlated with copy number in duplicated regions (Alkan et al. 2009). Excluding X, Y and random chromosomes, we predicted 112 Mbp of duplication in Kwanza and 113 Mbp in Kamilah, of which 102.7 Mbp was shared.

By way of comparison, we used data for a set of six HapMap individuals (NA18507, NA12878, NA12981, NA12892, NA19238, NA19239) (Sudmant et al. 2010), taking all 15 possible pairwise combinations and looking as before for private duplications bigger than 20 kbp in each pair. Fig. SF13.1 shows that the level of private duplication found in the two gorillas lies outside the range of human private duplications found in this way.

To get a more accurate measure of gorilla genomic duplication, we reclassified the WSSD duplication maps using array comparative genomic hybridization. Custom microarrays were created based on a customized oligonucleotide microarray (NimbleGen, 385,000 isothermal probes) targeted specifically to the gorilla segmental duplications. We then carried out a hybridization experiment using genomic DNA from Kwanza and Kamilah (the same samples as were used in sequencing), including replication with test and reference labels swapped. The log2 relative hybridization intensity was calculated for each probe, and we restricted our analysis to regions greater than 20 kbp in length and containing at least 20 probes. After normalization, specific duplications were detected if the median log2 of the region was beyond previously established thresholds (minimum 10 probes, median log2 of the region = 0.2) (Marques-Bonet et al. 2009).

Using this approach, we reclassified 289 duplications (totaling 10.1 Mbp), of which approximately 50% were individual-specific false positives, leaving 0.92 Mbp of validated duplications specific to Kwanza and 1.55 Mbp specific to Kamilah (Table ST13.1). By comparison, in a similar analysis of three humans (Alkan et al. 2009) only ~100 kbp of validated private duplications were found between any two individuals.

We also found 16 genes that are duplicated in one or other of Kamilah and Kwanza (6 completely and 10 partially duplicated) (Table ST13.2; Fig. SF13.2 shows an example at the SGMS1 locus).

Found in	WSSD duplications > 20 kbp (Mbp)	Validated duplications > 20 kbp (bp)	Number of validated duplications > 20 kbp
Kwanza only	3.1	923,398	24
Kamilah only	3.3	1,551,983	30
Kamilah and Kwanza	95	101,703,787	1359

Table ST13.1 Duplications longer than 20 kbp in Kwanza and Kamilah detected by WSSD and validated and reclassified by aCGH.

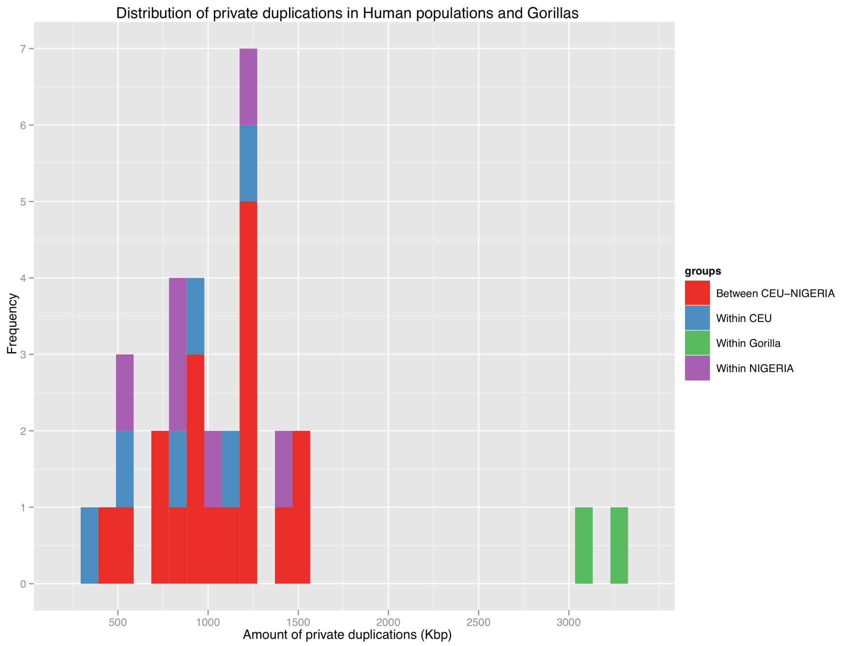


Figure SF13.1 Distribution of private duplications in 6 HapMap individuals (NA18507, NA12878, NA12981, NA12892, NA19238, NA19239) and two Western gorillas.

Gene ID	Protein ID	Gene name	Description	Specific dup	Log2 Kwanza/Kamilah aCGH	Complete coding
NM_001005513	NP_001005513	OR4C45	olfactory receptor, family 4, subfamily C	Kamilah specific	-0.63475	Complete
NM_001005512	NP_001005512	OR4A47	olfactory receptor, family 4, subfamily A	Kamilah specific	-0.8365	Complete
NM_001004703	NP_001004703	OR4C46	olfactory receptor, family 4, subfamily C	Kamilah specific	-0.278	Complete
NM_003739	NP_003730	AKR1C3	aldoketo reductase family 1, member C3	Kamilah specific	-0.23025	Partial
NM_147156	NP_671512	SGMS1	sphingomyelin synthase 1	Kamilah specific	-0.642	Partial
NM_001005242	NP_001005242	PKP2	plakophilin 2 isoform 2a	Kamilah specific	-0.3565	Partial
NM_177550	NP_808218	SLC13A5	solute carrier family 13, member 5 isoform a	Kamilah specific	-0.36975	Partial
NM_023073	NP_075561	C5orf42	hypothetical protein LOC65250	Kamilah specific	-0.32625	Partial
NM_000772	NP_000763	CYP2C18	cytochrome P450 family 2 subfamily C polypeptide	Kwanza specific	0.207	Partial
NM_152309	NP_689522	PIK3AP1	phosphoinositide3kinase adaptor protein 1	Kwanza specific	0.65425	Partial
NM_001005171	NP_001005171	OR52K1	olfactory receptor, family 52, subfamily K,	Kwanza specific	0.221	Complete
NM_144705	NP_653306	TEKT4	tektin 4	Kwanza specific	0.709	Complete
NM_133437	NP_597681	TTN	titin isoform novex2	Kwanza specific	0.275	Partial
NM_052958	NP_443190	C8orf34	hypothetical protein LOC116328	Kwanza specific	0.204	Partial
NM_002195	NP_002186	INSL4	insulinlike 4 precursor	Kwanza specific	0.5535	Complete
NM_000718	NP_000709	CACNA1B	calcium channel, voltagedependent, N type	Kwanza specific	0.3205	Partial

Table ST13.2: Gene table of gorilla specific validated duplications.

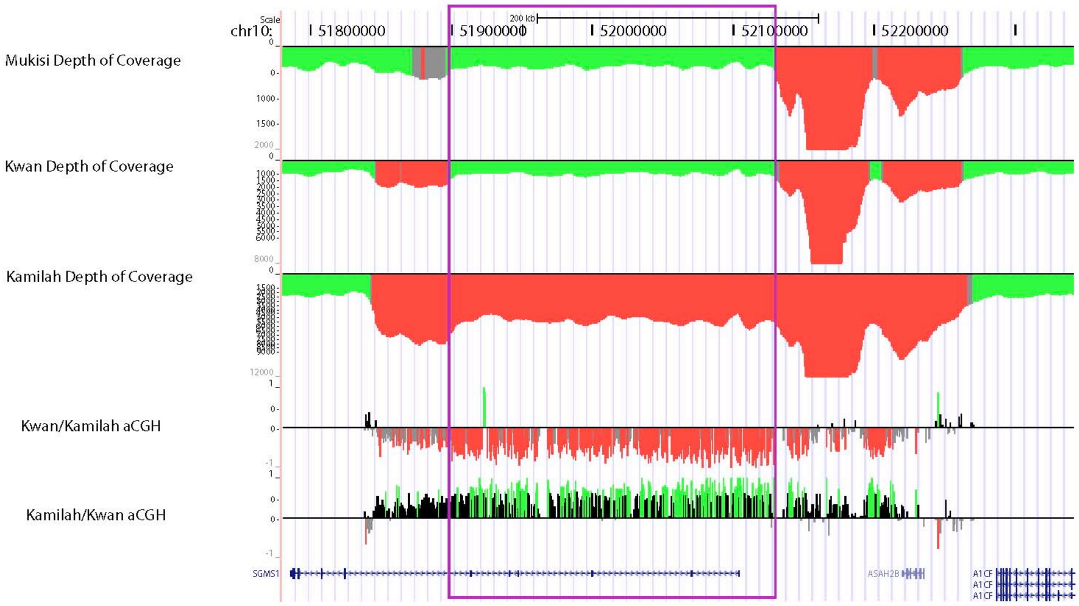


Figure SF13.2: An example of a validated gene (SGMS1) containing duplication in Kamilah. The top 3 tracks display the depth of coverage of whole-genome sequence data from gorillas Mukisi, Kwanza and Kamilah. In red, duplications are shown, when the depth of coverage exceeds 3 standard deviations of control regions. The purple square shows the Kamilah specific duplication. At the bottom, the results of the arrayCGH between Kwanza and Kamilah (and corresponding DyeSwap) and the gene track.

14. References

1000 Genomes Project Consortium (2010). "A map of human genome variation from population-scale sequencing." *Nature* **467**(7319): 1061-1073.

Alexa, A., J. Rahnenfuhrer, et al. (2006). "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure." *Bioinformatics* **22**(13): 1600-1607.

Awadalla, P., J. Gauthier, R. A. Myers, F. Casals, F. F. Hamdan, A. R. Griffing, M. Côté, E. Henrion, D. Spiegelman, J. Tarabeux, A. Piton, Y. Yang, A. Boyko, C. Bustamante, L. Xiong, J. L. Rapoport, A. M. Addington, J. L. DeLisi, M.-O. O. Krebs, R. Joober, B. Millet, E. Fombonne, L. Mottron, M. Zilversmit, J. Keebler, H. Daoud, C. Marineau, M.-H. H. Roy-Gagnon, M.-P. P. Dubé, A. Eyre-Walker, P. Drapeau, E. A. Stone, R. G. Lafrenière, and G. A. Rouleau (2010). 'Direct measure of the de novo mutation rate in autism and schizophrenia cohorts'. *American journal of human genetics.* **87**(3): 316-324

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. (2002) "Recent segmental duplications in the human genome". *Science*. **297**(5583):1003-7.

Bentley, D. et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry". *Nature*. **456**(7218):53-9.

Benjamin , Y. and Y. Hochberg (1995). "Controlling the Fals Discovery Rate: a Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* **57**(1): 289-300.

Bielawski JP, Dunn KA, Yang Z. (2000). Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics*. **156**(3):1299-308.

Bierne N, Eyre-Walker A. (2003).The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics*. **165**(3):1587-97.

Boehler C, Gauthier LR, Mortusewicz O, Biard DS, Saliou JM, Bresson A, Sanglier-Cianferani S, Smith S, Schreiber V, Boussin F, Dantzer F. (2011). Poly(ADP-ribose) polymerase 3 (PARP3), a newcomer in cellular response to DNA damage and mitotic progression. *Proc Natl Acad Sci*. 108(7):2783-8.

Burgess R, Yang Z. (2008) Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol.; **25**(9):1979-94.

Charnov, E., (2004) Evolutionary Ecology Research, **6**: 307–313

Chimpanzee Sequencing and Analysis Consortium (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome." *Nature* **437**(7055): 69-87.

Coyne, J. A. and Orr, H. A. (2004). *Speciation.* Sinauer Associates.

Down, T. A. and T. J. Hubbard (2005). "NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence." *Nucleic Acids Research* **33**(5): 1445-1453.

Dutheil, J. Y., G. Ganapathy, et al. (2009). "Ancestral population genomics: the coalescent hidden Markov model approach." *Genetics* **183**(1): 259-274.

Ebana Y, Ozaki K, Inoue K, Sato H, Iida A, Lwin H, Saito S, Mizuno H, Takahashi A, Nakamura T, Miyamoto Y, Ikegawa S, Odashiro K, Nobuyoshi M, Kamatani N, Hori M, Isobe M, Nakamura Y, Tanaka T. (2007). A functional SNP in ITIH3 is associated with susceptibility to myocardial infarction. *J Hum Genet*.;52(3):220-9.

Edvardson S, Jalas C, Shaag A, Zenvirt S, Landau C, Lerer I, Elpeleg O. (2011). A deleterious mutation in the LOXHD1 gene causes autosomal recessive hearing loss in Ashkenazi Jews. *Am J Med Genet A.* 155A(5):1170-2.

Ellegren, H. (2009). "A selection model of molecular evolution incorporating the effective population size." *Evolution* **63**(2): 301-305.

Ewens, W. J. (2004). *Mathematical population genetics*, Springer.

Fleagle, J. G. (1998). *Primate Adaptation and Evolution, Second Edition*. Academic Press.

Flicek, P., M. R. Amode, et al. (2011). "Ensembl 2011." *Nucleic Acids Research* **39**(Database issue): D800-806.

Gibbs, R. A., J. Rogers, et al. (2007). "Evolutionary and biomedical insights from the rhesus macaque genome." *Science* **316**(5822): 222-234.

Gillooly, J. F., Allen, A. P., West, G. B., Brown, J. H. (2005). "The rate of DNA evolution: Effects of body size and temperature on the molecular clock". *PNAS* **102**(1): 140-145.

Goodman, M. (1961). 'The role of immunochemical differences in the phyletic development of human behavior'. *Hum. Biol.* **33**, 131-162.

Hein, J., M. H. Schierup, et al. (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory,* Oxford University Press.

Hobolth, A., O. F. Christensen, et al. (2007). "Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model." *PLoS genetics* **3**(2): e7.

Hobolth, A., J. Y. Dutheil, et al. (2011). "Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection." *Genome Research* **21**(3): 349-356.

Kasowski, M., F. Grubert, et al. (2010). "Variation in transcription factor binding among humans." *Science* **328**(5975): 232-235.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." *Genome Research* **12**(4): 656-664.

Kim S. H., Elango N., Warden C., Vigoda E., Yi S.V. 'Heterogeneous genomic molecular clocks in primates'. *PLoS Genet*. 2006. **2**(10):e163

Kondrashov, A. S. (2003). 'Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases'. *Hum. Mutat.* **21**(1): 12-27

Kong, L., Y. Zhang, et al. (2007). "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." *Nucleic Acids Research* **35**(Web Server issue): W345-349.

Kosiol, C., T. Vinar, et al. (2008). "Patterns of positive selection in six Mammalian genomes." *PLoS genetics* **4**(8): e1000144.

Langmead, B., K. D. Hansen, et al. (2010). "Cloud-scale RNA-sequencing differential expression analysis with Myrna." *Genome biology* **11**(8): R83.

Li, H. and R. Durbin (2010). "Fast and accurate long-read alignment with Burrows‚ÄìWheeler transform." *Bioinformatics* **26**(5): 589-595.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* **25**(16): 2078-2079.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." *Genome research* **18**(11): 1851-1858.

Lindblad-Toh, K., C. M. Wade, et al. (2005). "Genome sequence, comparative analysis and haplotype structure of the domestic dog." *Nature* **438**(7069): 803-819.

Lynch, M. (2010). 'Rate, molecular spectrum, and consequences of human mutation'. *Proceedings of the National Academy of Sciences* **107**(3): 961-968.

Massingham, T. and N. Goldman (2005). "Detecting amino acid sites under positive selection and purifying selection." *Genetics* **169**(3): 1753-1762.

Matsumura, S., Forster, P, (2008). Generation time and effective population size in Polar Eskimos. *Proc R Soc B* **275**:1501-1508

McConkey, E. H. (2004). "Orthologous numbering of great ape and human chromosomes is essential for comparative genomics." *Cytogenetic and genome research* **105**(1): 157-158.

McDaniell, R., B. K. Lee, et al. (2010). "Heritable individual-specific and allele-specific chromatin signatures in humans." *Science* **328**(5975): 235-239.

Meader, S., L. W. Hillier, et al. (2010). "Genome assembly quality: assessment and improvement using the neutral indel model." *Genome Res* **20**(5): 675-684.

Montgomery, S. B., M. Sammeth, et al. (2010). "Transcriptome genetics using second generation sequencing in a Caucasian population." *Nature* **464**(7289): 773-777.

Mullikin, J. and Z. Ning (2003). "The Phusion Assembler." *Genome Research* **13**(1): 81-90.

Nielsen, R. and J. Wakeley (2001). "Distinguishing migration from isolation: a Markov chain Monte Carlo approach." *Genetics* **158**(2): 885-896.

Nielsen, R. and Z. Yang (1998). "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene." *Genetics* **148**(3): 929-936.

Ning, Z., A. J. Cox, et al. (2001). "SSAHA: a fast search method for large DNA databases." *Genome research* **11**(10): 1725-1729.

Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. (2008). 'Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs'. *Genome Research* **18**, 1814-1828.

Pruitt, K. D., T. Tatusova, et al. (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic Acids Research* **33**(Database issue): D501-504.

R_Development_Core_Team. (2006). "R: A Language and Environment for Statistical Computing." from http://www.R-project.org.

Revol De Mendoza, A., D. Esquivel Escobedo, et al. (2004). "Expansion and divergence of the GH locus between spider monkey and chimpanzee." *Gene* **336**(2): 185-193.

Roach, J. C., G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas (2010). 'Analysis of genetic inheritance in a family quartet by Whole-Genome sequencing'. *Science* **328**(5978): 636-639.

Rouleau M, Saxena V, Rodrigue A, Paquet ER, Gagnon A, Hendzel MJ, Masson JY, Ekker M, Poirier GG. (2011). A key role for poly(ADP-ribose) polymerase 3 in ectodermal specification and neural crest development. *PLoS One*. **6**(1):e15834.

Sawyer, S. L., M. Emerman, et al. (2004). "Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G." *PLoS biology* **2**(9): E275.

Schmidt, D., M. D. Wilson, et al. (2009). "ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions." *Methods* **48**(3): 240-248.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. (2009). Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol*. **1**:114-8.

Simpson, J., K. Wong, et al. (2009). "ABySS: a parallel assembler for short read sequence data." *Genome research* **19**(6): 1117-1123.

Singleton, I., Wich, S.A., Griffiths, M. (2008). 'Pongo abelii'. In: *IUCN 2011. IUCN Red List of Threatened Species*. Version 2011.2.)

Skaletsky, H., T. Kuroda-Kawaguchi, et al. (2003). "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes." *Nature* **423**(6942): 825-837.

Slatkin M, Pollack JL. (2008). Subdivision in an ancestral species creates asymmetry in gene trees. *Mol Biol Evol*. **25**(10):2241-6.

Stanyon, R., M. Rocchi, et al. (2008). "Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres." *Chromosome research* **16**(1): 17-39.

Steiper, M., N. M. Young (2006), 'Primate molecular divergence dates'. *Molecular Phylogenetics and Evolution* **41**: 384-394.

Tamura, K., J. Dudley, et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." *Molecular Biology and Evolution* **24**(8): 1596-1599.

Teleki, G., E. E. Hunt Jr., and J. H. Pfiffering. (1976). Demographic observations (1963-1973) on the chimpanzees of Gombe National Park, Tanzania. *J. Hum. Evol*. **5**:559-598

Thalmann, O., A. Fischer, et al. (2007). "The complex evolutionary history of gorillas: insights from genomic data." *Mol Biol Evol* **24**(1): 146-158.

Vilella, A. J., J. Severin, et al. (2009). "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates." *Genome Res* **19**(2): 327-335.

Wallis, M. (2001). "Episodic evolution of protein hormones in mammals." *Journal of molecular evolution* **53**(1): 10-18.

Warren, W. C., L. W. Hillier, et al. (2008). "Genome analysis of the platypus reveals unique signatures of evolution." *Nature* **453**(7192): 175-183.

Wood, B. and T. Harrison (2011). "The evolutionary context of the first hominins." *Nature* **470**(7334): 347-352.

Wyckoff, G. J., W. Wang, et al. (2000). "Rapid evolution of male reproductive genes in the descent of man." *Nature* **403**(6767): 304-309.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." *Molecular biology and evolution* **24**(8): 1586-1591.

Yang Z, Nielsen R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*. **46**(4):409-18.

Yang, Z. and R. Nielsen (2002). "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages." *Molecular Biology and Evolution* **19**(6): 908-917.

Young, M. D., Matthew J. Wakefield, Gordon K. Smyth, Alicia Oshlack (2010). *Genome Biology* 2010, 11:R14, 2010

Zerbino, D. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." *Genome Research* **18**(5): 821-829.

Zhang, J., R. Nielsen, et al. (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." *Molecular Biology and Evolution* **22**(12): 2472-2479.

Zhang, L., Lu, H. H. S., Chung, W., Yang, J. Li, W-H. (2004). 'Patterns of Segmental Duplication in the Human Genome'. *Mol Biol Evol* **22**(1): 135-141

Zhang, Y., T. Liu, et al. (2008). "Model-based analysis of ChIP-Seq (MACS)." *Genome biology* **9**(9): R137.

Zhang, Z., N. Carriero, et al. (2004). "Comparative analysis of processed pseudogenes in the mouse and human genomes." *Trends in genetics* **20**(2): 62-67.